University of Glasgow

Towards a portable mid-infrared tool for analysis of mosquito vectors of malaria

Mauro Pazmino Betancourth

Submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

Contents

List of Tables

List of Figures

Abstract

Mid-infrared spectroscopy (MIRS) has emerged as a potential tool to predict species, age and infection in the *Anopheles malaria mosquitoes* as well as in other disease vectors. The main advantages of optical methods in general are their speed, little or no sample preparation, label-free, lower cost and already established protocols and analysis pipelines. New rapid, low cost, high-throughput tools for vector surveillance are urgently needed to develop and optimise new vector control strategies, as vector borne diseases (VBD) are spreading around the globe due to climate change and globalisation, and endemic countries are suffering resurgence of malaria cases following weakening of control tools. However, the current commercially available FTIR spectrometers have limitations. They are expensive, bulky and low power that hider its implementation in the field. Quantum cascade lasers (QCL) have become an alternative to FTIR light sources due to their unique characteristics (i.e. coherence, high power in a smaller spot size, small chip size), which allows easier implementation for the field due to its lower cost, practicality, and accuracy. These characteristics can expand the possibilities to develop new ways to measure spectral information from disease vectors. This thesis is aimed at developing a QCL-based spectrometer, understanding the most informative infrared region for VBD surveillance and the use of legs for surveillance and prediction of key traits from mosquitoes using MIRS.

In this project, micro-diffuse reflectance spectroscopy (μDRIFT) was used on mosquito legs to predict age, species and cuticular insecticide resistance. Indeed, spectra from legs led to high accuracy ML models for age prediction (overall model accuracy: 77.1% (± 6.5%) with a mean accuracy of 82% for 3 days old and 74% for 10 days old) and moderate accuracy for species identification (overall model accuracy: 69.1% (± 7.9%) with a mean accuracy of 68% for *An. gambiae* and 71% for *An. coluzzii*). Finally, cuticular resistance in three strains of *Anopheles* mosquitoes was identified with high accuracy when grouped into susceptible and resistant classes (overall model accuracy: 71.3% (± 8%) with a mean accuracy of 73% for susceptible and 71% for resistant class). However, these preliminary

findings need to further be confirming by ruling out confounding factors such as the use of different strains of *Anopheles* by using a single strain with various degrees of insecticide resistant. I found that Partial Least Squares Discriminant Analysis (PLS-DA) and can be used for high accuracy prediction between *An. gambiae* and *An. coluzzii* when tested on laboratory samples from the same origin (mean accuracy: 87%). However, species prediction decreases when the model is tested on samples from different laboratories (mean accuracy: 62%) and in semi-field samples (mean accuracy: 46.5%). For age prediction, PLS regression was able to predict different group ages (3, 5, 7, 9, 12, 15 days old) when tested with laboratory samples from the same origin (R^2 = 0.68, RMSE = 2.24) and with samples from other laboratories (R^2 = 0.78, RMSE = 1.89). Nevertheless, the model cannot predict the age of semi-field samples (R^2= -1.84, RMSE = 7.99). Also, I found narrower spectral windows of $\approx$ 300 cm^{-1} in length located on the Amide I and Amide II region are sufficient to predict mosquito species using machine learning (accuracy from 88% to 98%). This can help for a more efficient way of collecting spectral data. Future work should focus on how to improve model calibration by adding samples with diverse origin (different laboratories, different rearing conditions) to improve model generalisation. Finally, I have developed a QCL-based spectrometer in the range of 8-11 μm with scan speeds up to 500 Hz, with a maximum tuning rate of 400 μm/s. The system can collect spectra from polymers (polypropylene, polyethylene terephthalate and polyethylene) and biological samples (mosquitoes) in transmission mode. When compared to commercial FTIRs, MIRS measurements of whole mosquito bodies in KBr discs through the QCL-based spectrometer were in high agreement at bands 988, 1029 and 1056 cm^{-1} showing that the newly developed device works in mosquitoes. This study has made the first step towards the use of QCL-based system for spectroscopy of insect disease vectors, opening new opportunities for the implementation and use of mid-infrared spectroscopy for vector-borne disease surveillance.

Declaration of Authorship

I declare that this thesis is the result of my own work, except where explicit reference is made to the work of others, and has not been presented in any previous application for a degree at this or any other institution.

Publications/Presentations

Conference Papers

M. Pazmino-Betancourth, V. Ochoa-Gutierrez, and D. Childs, "Organic polymers classification using QCL spectroscopy," in *OSA High-brightness Sources and Light-driven Interactions Congress 2020 (EUVXRAY, HILAS, MICS)*, L. Assoufid, P. Naulleau, M. Couprie, T. Ishikawa, J. Rocca, C. Haefner, G. Sansone, T. Metzger, F. Quéré, M. Ebrahim-Zadeh, A. Helmy, F. Laurell, and G. Leo, eds., OSA Technical Digest (Optical Society of America, 2020), paper MF1C.6.

Papers to be submitted in scientific journals

Pazmino, M., Baldini, F., Wynne K., Hogg, R., Childs, D. "Diffuse reflectance spectroscopy for species, age and cuticular resistance prediction in *Anopheles gambiae*"'. To be submitted to Malaria Journal or Applied Spectroscopy.

Pazmino, M., Boldin A., Ochoa-Gutierrez, V., Hogg, R., Childs, D. 2021. "Portable External Cavity Quantum Cascade for mid infrared spectroscopy applications". To be submitted to Optic Letters.

Oral Presentations

Pazmino, M. 2021. "Quantum Cascade Lasers for Mosquito Surveillance". ASTMH Annual Meeting. Online

Pazmino, M., Boldin A., Ochoa-Gutierrez, V., Hogg, R., Childs, D. 2021. "Portable External Cavity Quantum Cascade for mid infrared spectroscopy applications". Interna-

tional Conference on Solid State Devices and Materials (SSDM2021). Online

Pazmino, M, Ochoa-Gutierrez, V. Childs, D. 2020. "Organic polymers classification using QCL spectroscopy". Mid-Infrared Coherent Sources Optical Society of America. Online

Pazmino, M. 2020. "Identification of Malaria Vectors and insecticide resistant using Mid-infrared Spectroscopy through a portable Quantum Cascade Laser device". ASTMH Annual Meeting. Online

Pazmino, M, 2019. "Diffuse Reflectance for assessing mosquito population structure". Oral presentation. International Symposium & Annual National Science Meeting. 'Vectors of diseases'. London, UK

Pazmino, M. 2019. "Tuneable Quantum Cascade Lasers for the Mapping of Mosquito Population Structure". UK Semiconductor Conference. Sheffield, UK.

Poster Presentations

Pazmino, M. 2021. "Quantum Cascade Lasers for Mosquito Surveillance". ASTMH Annual Meeting. Online

Pazmino, M.,2021. "Identification of malaria vectors and insecticide resistant using micro diffuse reflectance spectroscopy and quantum cascade lasers". PAMCA Annual conference. Online

Pazmino, M.,2020. "Identification of malaria vectors and insecticide resistant using mid-infrared spectroscopy through a portable quantum cascade laser device". PAMCA-VectorBase Virtual Vector Meeting. Online

Pazmino, M. 2018. "Tuneable Quantum Cascade Lasers for the Mapping of Mosquito Population Structure". 8th International Quantum Cascade Laser School and Workshop. Cassis, France

Publications from collaborations

V. Ochoa-Gutierrez, **M. Pazmino-Betancourth**, J. Reboud, A. R. Harvey, and J. M. Cooper, "Analysis and exploration of heme groups using ATR-FTIR for future health

monitoring" in Proc.SPIE (2021), Vol. 11879.

R. I. Soria, S. A. Rolfe, **M. P. Betancourth**, and S. F. Thornton, "The relationship between properties of plant-based biochars and sorption of Cd(II), Pb(II) and Zn(II) in soil model systems," Heliyon 6, e05388 (2020).

Acknowledgements

I would like to thank my wife, Mishell, for all the love through the good, the bad and the ugly of these four years, I love you. To my mum, Monica, who taught me how not to be afraid of learning new things. To my sisters, Tefa for all her support and Beatriz for been my academic mentor and role model, without her guidance, I wouldn't been here.

I would like to thank all the new and old friends I made during my stay in Glasgow: Leonardo for keeping me doing mosquito research. Victoria, for all these years of friendship, and especially to be there during the COVID-19 lock-downs. Victor for pushing me to do science for the people. To Jafet, The Mexican crew, Delos, Suzan, Vale, the LKAS colleagues and the people I've met these four years.

This research wouldn't been possible without the immense help I got from many people. First, Aleksadr Boldin and his incredible knowledge about electronics and for all his help with the laser set up and programming. All the people from the photonic devices and systems group (Aye, Danqi, Zijun, Razvan). Dr. Mario Gonzalez at the Wynne Lab for his help with the FTIR, spectroscopy measurements and overall advice about spectroscopy. Dr. Matthew Steer for its invaluable advice on characterisation of the laser. To the Lord Kelvin Adam Smith scholarship for funding the project.

And finally, I would like to express my sincere gratitude to my supervisors, Dr. David Childs for his guidance and patient and Dr. Francesco Baldini for his mentorship, support, and especially his trust on this project. Dr. Heather Ferguson for her insightful comments on this thesis. Dr. Klaas Wynne, Dr. Richard Hoggs for their support completing the final thesis.

List of Abbreviations

AlInAs	Aluminium indium arsenide
AlN	aluminum nitride
AlSb	Aluminium indium arsenide
ATR	Attenuated Total Reflection
CW	Continuous wave
EC-QCL	External Cavity Quantum Cascade Laser
EIP	Extrinsic incubation period
FTIR	Fourier Transform infrared
FWHM	Full width at half maximum
GaAs	Gallium arsenide
GaInAs	Gallium Indium arsenide
InAs	Indium arsenide
InP	Indium Phosphide
IRS	Insecticide residual spray
ITN	Insecticide treated net
MCT	Mercury cadmium telluride
MOEMS	Microoptoelectromechanical systems
MIRS	Mid infrared spectroscopy
NIRS	Near infrared spectroscopy
PCA	Principal component analysis
PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
µDRIFT	microdiffuse reflectance spectroscopy

Chapter 1

Introduction

Malaria is one of the most important vector-borne diseases, with 241 million cases and 627 000 deaths recorded in 2020 worldwide [1]. It is mainly transmitted by 4 major vectors: *Anopheles gambiae, An. coluzzii, An. arabiensis* and *An. funestus* [2]. However, other species such as *An. rivulorum*, *An. vaneedeni*, *An. leesoni* and *An. merus* can act as secondary vectors [3–5]. Malaria cycle starts with the mosquito ingesting the parasite through a blood meal from an infected host. The parasite requires from 9 to 14 days to develop into its infectious stage and to move to the salivary glands, where it is transmitted to the next host when the mosquito blood feeds again [6, 7]. Thus, the mosquito must survive long enough for the parasite to be successfully transmitted. Vector control is the most effective way of reducing malaria transmission [8]. The target of current vector control approaches is to decrease the life span of adult mosquitoes by using Long Lasting Insecticidal Nets (LLINs), and Insecticide Residual Spray (IRS) [9]. However, mosquito populations are becoming increasingly resistance to insecticides across Africa [10, 11].

Vector surveillance is the core of vector control programmes in endemic countries [12, 13]. The aim of surveillance is to estimate the intensity of transmission and assess the impact and progress of intervention programs [14]. Surveillance includes estimating species composition, abundance, infection rate of vector populations, and insecticide resistance status [15]. In addition, other entomological parameters, such as the age structure of a population, are valuable to assess due to its influence on transmission [16]. Existing methods for age grading, species identification and insecticide resistance can be labour-intensive, expensive, and in the case of age grading, not accurate. In the search of cheaper, accurate and faster methods, infrared spectroscopy has become a potential tool for mosquito surveillance [17–19]. The ability to predict species, age, and insecticide resistance by detecting chemical changes in the mosquito's cuticle using infrared

spectroscopy opens an opportunity to improve current surveillance systems. Furthermore, the use of better and faster machine learning algorithms to analyse infrared data promises an accurate and versatile tool in the fight against vector-borne diseases. Thus, the project "Towards Portable Mid-Infrared Tool for Analysis of Mosquito Vectors of Malaria" aims at improving vector surveillance by exploring different alternatives to current spectroscopy methods and the use of next generation semiconductor lasers as a tool for mosquito surveillance.

This chapter will provide an overview of the current state-of-the-art infrared spectroscopy. It will start describing the fundamentals of spectroscopy, the different sampling techniques with emphasis in the ones used in the following experimental chapters. Then, it will explain how quantum cascade works and its current applications. Also, there will be an overview on machine learning, and which are the most common techniques for species identification, age grading and identification of insecticide resistance status in mosquitoes. It will finish with a review of the current state of infrared spectroscopy applied in mosquito surveillance.

1.1 Infrared Spectroscopy

Infrared spectroscopy is one of the most important and widely used analytical techniques [20]. The main advantage is its versatility, as any sample from films, solutions, powders, films, fibre types, etc. can be studied with spectroscopy [21]. Infrared spectroscopy (near infrared, mid-infrared, and Terahertz) along with Raman spectroscopy are techniques of vibrational spectroscopy [22]. Infrared spectroscopy measures the transitions between molecular vibrational energy levels, which are the result of mid-infrared absorption. The characteristics of infrared vibrational bands are given by their frequency, intensity, and band shape [23]. The vibrational energy levels are unique to each molecule; thus, infrared spectroscopy can provide the fingerprint of absorbed frequencies of a molecule of interest [24]. The measurement of these absorbed frequencies is presented in the form of an infrared spectrum [20]. Therefore, by applying this technique to a sample, it will generate a specific infrared spectrum base on the sample's chemical composition. Applications of infrared spectroscopy can be seen in agriculture, food science, chemistry, security, and biomedical sciences [24–26].

1.1.1 Fundamentals of electromagnetic radiation

Light is electromagnetic radiation which propagates through space as a wave [22]. It consists of an electric field and a magnetic field, which oscillate in single planes perpendicular to each other [20]. They are described by a continuous sinusoidal wave-like motion. Infrared spectroscopy neglects the magnetic field component and only consider the electrical field component [23] as shown in Figure 1.1.

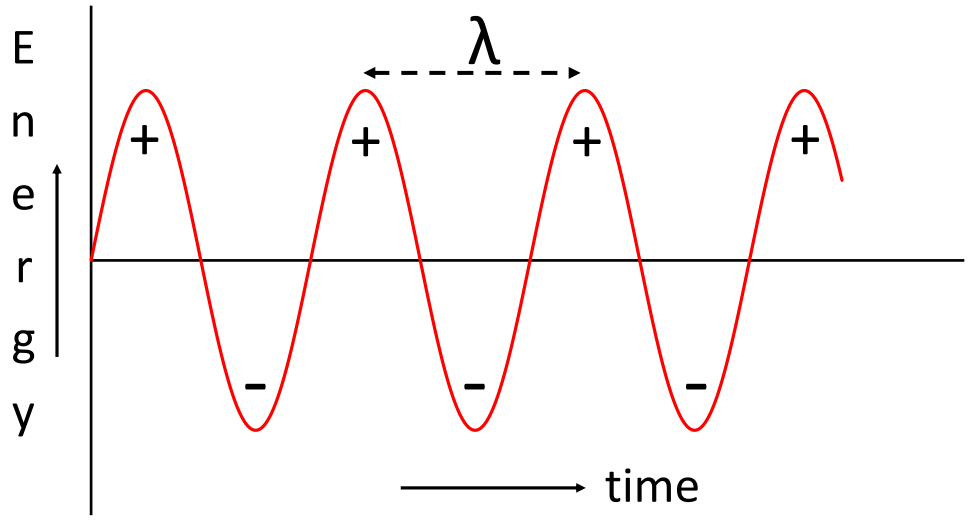


Figure 1.1: Amplitude of the electric field component of the electromagnetic radiation as a function of time. Wavelength (λ) is the distance between two crests (Plot adapted from Larkin, P. 2017 [23]).

The velocity of the propagation of electromagnetic radiation is constant, which is the speed of light (c). In a fixed distance, the speed of the wave is the product of wavelength (λ) which is the distance between two crests and the frequency (ν), defined as the number is cycles per unit of time (seconds) as show in the following equation:

$$c = \lambda * (\nu) \tag{1.1}$$

Wavelength units are represented as wavenumbers, $\tilde{\nu}$, in cm^{-1} which is defined as the number of waves per unit of length (one centimetre) [20, 23]. These parameters are related to the following equation:

$$\tilde{\nu} = \frac{\nu}{c} \tag{1.2}$$

We picture a beam of light as a wave that contains a certain amount of radiant energy,

which can be subdivided into a certain number of photons. The amount of energy in one photon is given by:

$$E = h\nu \tag{1.3}$$

Here, h is the Planck constant and ν is the photon frequency [27].

1.1.2 Infrared Absorption

A molecule when is irradiated with infrared light, it can absorb the light under specific conditions. First, the energy $h\upsilon$ of the infrared light absorbed by the molecule must be equal to the energy difference between two energy levels of vibration of the molecule (ground state (E_0) to excited state (E_1)) [20, 22, 23, 28] (Fig. 1.2). This phenomenon can be represented by the following equation:

$$\nu = \frac{(E_1 - E_0)}{h} \tag{1.4}$$

Equation 1.4 is called the Bohr frequency condition. Therefore, infrared absorption is based on a transition between energy levels of a molecular vibration [22]

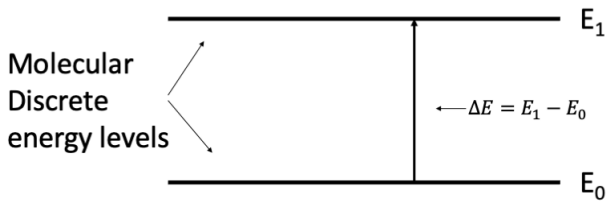


Figure 1.2: Illustration of discrete energy levels and absorption of electromagnetic radiation (Plot adapted from Larkin, P., 2017 [23]).

The selection rule for infrared radiation is that transitions with a change in the vibrational quantum number by ± 1 are allowed, while other transitions are forbidden. Another selection rule is that infrared light is absorbed when the electric dipole moment of a molecule changes. This will be defined by the symmetry of the molecule [22, 29].

Normal modes of vibration

When a molecule absorbs infrared light, its bonds vibrate. The vibrational motion caused by infrared light is complex, but it can be explained by simple vibrations called normal

modes. Normal modes can be described as independent motion of atoms in a molecule [30]. In each normal vibration, all atoms vibrate with the same frequency (normal frequency), and they pass through their equilibrium positions simultaneously. To calculate the number of normal modes of a molecule with N atoms, we use the rule 3N-5 and 3N-6 for linear and non-linear molecules, respectively. A diatomic molecule will have one normal mode, stretching vibration, where the bond can stretch and compress. Examples of these molecules are H_2, N_2 and O_2. Let's consider a three-atom non-linear molecule such as water (H_2O), it will have three normal modes. One symmetric stretch, one asymmetric stretch and one bending vibration as shown in Fig. 1.3a. For a linear molecule such as CO_2, it will have 4 normal modes. These modes are an asymmetric stretch, symmetric stretch and two bending vibrations (Fig. 1.3b). The number of normal modes is just a guide of how many infrared bands a molecule has, since not all vibrations can be excited by infrared light.

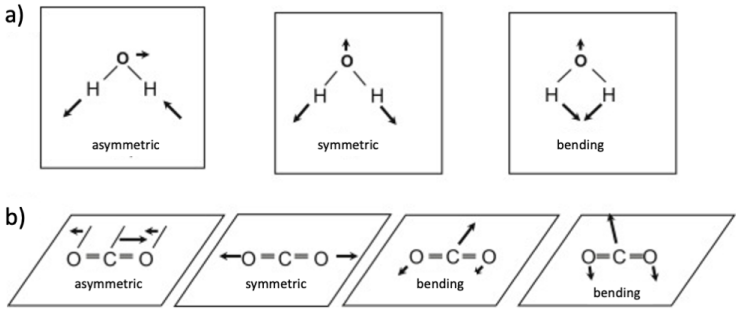


Figure 1.3: Representation of the normal modes of a linear molecule, carbon dioxide, and a non-linear molecule water (Adapted from Larkin, P. 2017 [23]).

Infrared absorption process

There are two conditions that need to be meet for infrared absorption. The first one is the change in dipole. For this, I am going to use the HCl molecule as a model. The HCl molecule is represented as a harmonic oscillator, where the atoms are represented as balls with masses and the bond is represented as a spring (Fig. 1.4a). The electronegativity difference between the two atoms results in the chlorine partially negative charge (δ^-) and the hydrogen partially positive charge charged (δ^+). The molecule has two charges separate by a distance; this is called a dipole moment, which is the measurement of the charge asymmetry of a molecule. It can be calculated by the equation:

$$\mu = q * r \tag{1.5}$$

where q is the charge, r is the distance and μ is the magnitude of the dipole. As mentioned before, the polarity of the electrical vector of electromagnetic radiation alternates from positive to negative through time, as shown in Fig. 1.1. When the electrical vector encounters the HCl molecule, depending on the polarity of the electrical vector, it will be repelled or attract the H atom, by doing so, the HCl bond will be shortened or stretched. The length of the bond will change with the same frequency as the polarity of the electric vector changes (Fig.1.4b). The molecule is vibrating with the same frequency as the electric vector; therefore, the energy of the photon is transferred to the molecule. We can represent this condition by the following equation:

$$\partial\mu/\partial X \neq 0 \tag{1.6}$$

where $\partial\mu$ is the change in the dipole moment and ∂X is the change in the bond distance.

"Infrared active" molecules will have a vibration that satisfies equation 1.6, the electric dipole moment changes as its bonds contract or expand. This is seen with heteronuclear diatomic molecules (molecules with two non-identical atoms). On the other hand, an "infrared inactive" molecule, its electric dipole moment remains zero no matter how much its bonds expand. This is the case of mononuclear diatomic molecules [31, 32].

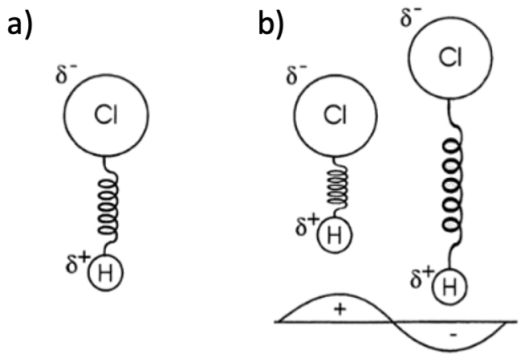


Figure 1.4: a) Diagram of a diatomic atom, HCl as harmonic oscillator. δ^+ and δ^- represent the partial charges of chlorine and hydrogen. b) Representation of the interaction between the HCl molecule and the electric vector. The change in polarity of the electric vector changes the length of the HCl bond (Modified from Smith, B. 1999 [33])

The potential energy of the classical harmonic oscillation of a diatomic atom is given by $PE = 1/2KX^2$, where X is the distance between masses. The plot of the potential energy as a function of X is a symmetric parabola, as shown in Fig. 1.5a. X_e is the

equilibrium bond distance where the energy is at minimum and K is the force constant measure of the curvature of the potential near X_e. According to quantum mechanics, molecules can only exist in quantise energy states, thus vibrational energy can only have discrete values. The vibrational energy levels are shown in Fig.1.5b. In this model, energy states are equidistant. However, a more realistic approach to the harmonic oscillator is anharmonicity, which happens if the change in dipole moment is not linearly proportional to the nuclear displacement. Fig.1.5b shows how the potential energy of a water molecule measure in wavenumber (y-axis) is affected by the changes in the bond distance measure in Angstrom (x-axis). As the bond distance is made shorter or longer than the equilibrium bond distance, the molecular energy increases. The discrete vibrational energy levels are notated with vibrational quantum number ν. The ground state of the molecule is $\nu = 0$, the next energy level, $\nu = 1$, is the first excited vibrational state. The separation between levels is not equidistant, and they become smaller at higher vibrational levels. In the case of H-O stretching vibration, for example, the distance between $\nu = 0$ and $\nu = 1$ is 3500 cm^{-1}. In order to a photon be absorbed by the molecule, it needs to match the energy difference between $\nu = 0$ and $\nu = 1$. Water absorbs photons with 3500 cm^{-1}. Therefore, the second condition for infrared absorption is that the energy of the photon needs to match a vibrational energy difference within a molecule [23,33,34].

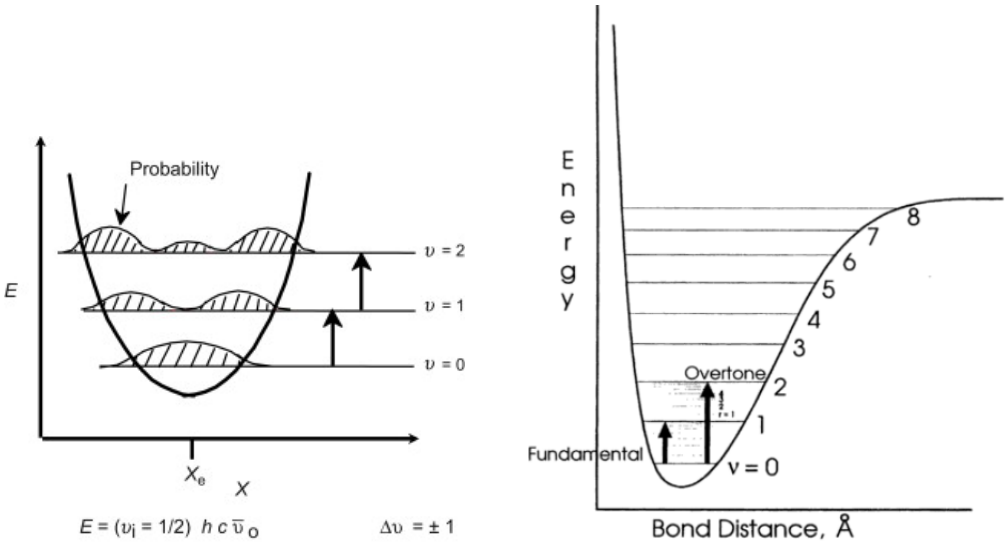


Figure 1.5: a) Potential energy E as a function of bond distance X for a diatomic molecule using the quantum harmonic oscillator model. $\nu = 0$, $\nu = 1$, $\nu = 3$ are vibrational energy levels b) Potential energy as a function of bond distance for water molecule using anharmonicity. Adapted and modified from Larkin, P. 2017 [23] and Smith, B. 2018 [31]

Group Frequencies

Group frequencies can help with the analysis of vibration spectra of polyatomic molecules. The concept can be applied when the amplitudes of a specific functional group are very large compared to other atoms. Examples of groups frequencies are stretching vibration of O-H group, C=O vibration of a carbonyl group, etc. These group frequencies are important in the analysis of infrared spectra [35].

Vibrations include a change in bond length, which is called stretching or a change in bond angle, called bending. Stretching can be in phase or symmetrical or out of phase or asymmetrical. For example, a molecule AX_2 will have six vibrational modes. Two of them stretching vibrations, one being symmetric stretching vibration and the other asymmetric stretching vibration. The rest are bending vibrations. This can be classified as scissoring, rocking, wagging, and twisting vibrations. Scissoring and rocking vibrations are in-plane vibrations, while wagging and twisting vibrations are out-of-plane vibrations (Fig. 1.6). Group frequencies become handy for more complex molecules such as proteins. A classic example is the amide group. The normal vibration of the amide group has been calculated using the N-methyl acetamide as a model. Out of the twelve normal modes of the molecule, three normal modes, amide I, II and III are key to study the structure of proteins. Amide I has a C=O stretching vibration. Amides II and III are coupling modes of C-N stretching vibrations and N-H in-plane bending vibrations, as shown in Fig. 1.7 [20, 30, 32]

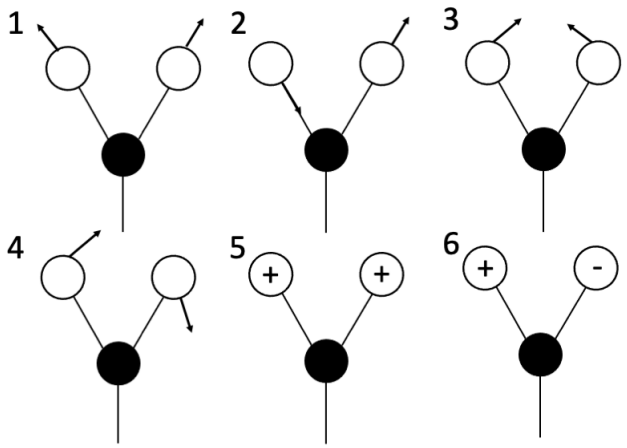


Figure 1.6: Vibrations of AX2 group. 1: symmetric stretching vibration. 2: antisymmetric stretching vibration. 3: scissoring vibration. 4: rocking vibration. 5: wagging vibration. 6: twisting vibration. Plot taken from Ozaki, Y. et al. 2021 [30]

Figure 1.7: Three normal vibrations of N-methylacetamide (model of an amide group). Plot taken from Ozaki, Y. et al. 2021 [30]

The intensity of the absorption of infrared light will be influenced by how greater the change of dipole is and how many infrared active vibrations a molecule has. For example, The number of infrared active vibrations will be less in symmetrical molecules than asymmetrical ones, which will be representing in weaker absorption bands. The vibrations and their characteristic absorption wavenumber are indicated in table 1.1.

Table 1.1: Vibrations and their characteristic absorption wavenumbers (Table modified from Smith, B. 1999 [31])

Group vibration	Classification	Wavenumber cm^{-1}
OH stretch	Free	~3600
	H-bounded	3000-2500
CH stretch	Olefin	~3080
	Methyl	~2960, ~2870
	Methylene	~2925, ~2850
SH stretch	–	2600-2550
C=O stretch	Acyl chloride	~3080
	Ester	~1735
	Aliphatic Aldehyde	~1730
	Aliphatic ketone	~1715

Quantitative Analysis

In absorption spectroscopy, the spectrometer will emit broadband infrared light. Due to absorption of infrared light by molecules, the light that passes through the sample will

be attenuated [36]. It is assumed that the relationship between the intensities of incident and transmitted infrared radiation and the concentration of analyte material present in a sample follows Lambert-Beer law which states: "The transmission of a sample within an incident bean is equivalent to 10 exponent the negative product of the molar extinction coefficient multiplied by the concentration of a molecule in solution time the path length of the sample in solution" [37]:

$$T = \frac{I}{I_0} = 10^{\epsilon c l} \tag{1.7}$$

Where T is transmittance, I_0 is the intensity of incident light, I is the intensity of transmitted light, ϵ is referred to as the molar absorptivity, c is the concentration and l is the path length. In the case of where Lambert-Beer law holds, the relationship between absorbance and analyte concentration is linear [23]. Absorbance is equal to the difference between the logarithms of the intensity of I_0 and I:

$$A = -log(\frac{I}{I_0}) \tag{1.8}$$

The infrared spectrum is represented by plotting transmittance or absorbance vs. wavenumber (Fig. 1.8).

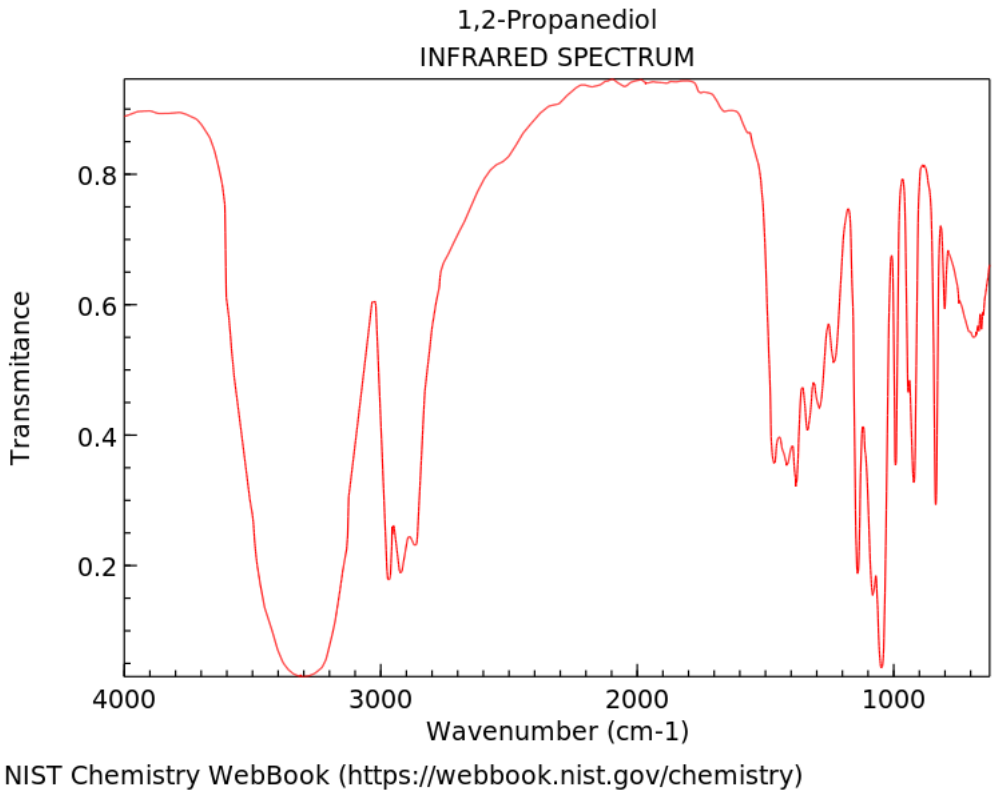


Figure 1.8: Example of an infrared spectrum from 1,2-Propanediol. Taken from National Institute of Standards and Technology

1.2 Spectrometers

Spectrometers are any instrument which measures a property of light as a function of wavelength [38]. The main components of a spectrometer are the light source, a detector, filters, mirrors, and gratings which will be assembled based on the type of spectrometer [37]. They can be classified into Dispersive and Fourier Transform. While disperse spectrometers were used extensively in the past, its limited resolution (fixed to one resolution width) and slow speed restricted its use [20]. Later, they were replaced by Fourier transform spectrometers, which currently dominated the area of infrared spectroscopy.

1.2.1 Disperse Spectrometer

Dispersive spectrometers produce spectra by taking incoming radiation from a source and dispersing into its spectral components. These types of spectrometers are also called grating or scanning spectrometers. The optical path of a dispersive spectrometer which uses a grating monochromator is shown in Fig. 1.9. They have three basic parts: a radiation source, a monochromator, and a detector. The incoming radiation goes through the entrance slit, and it is collimated onto the dispersive element. The dispersed radiation is then reflected back to the exit slit to the detector. The dispersive elements can include prisms and gratings. A grating consists of a reflecting surface with parallel grooves which are very closely spaced as shown in Fig. 1.10 right. Each groove diffracts the light to different directions, separating the light into its individual wavelengths. The reflected radiation is focused to the exit slit, and only radiation that leaves the grating at specific angles can exit through the slit. When radiation is reflected at an angle, the parallel beams from two adjacent grooves have a different path length (Fig. 1.10 right). The path length difference depends on the groove spacing, the angle of incidence (α), and the angle of reflectance of the radiation (β)(Fig. 1.10 right). This means that beams of a particular wavelength leaving at a specific angle from all the grooves will be in phase and show constructive interference when they converge at the exit slit, this is called the first order. Other wavelengths will show destructive interference. Different wavelengths will reach the detector by rotating the grating [20, 23, 34, 39–41].

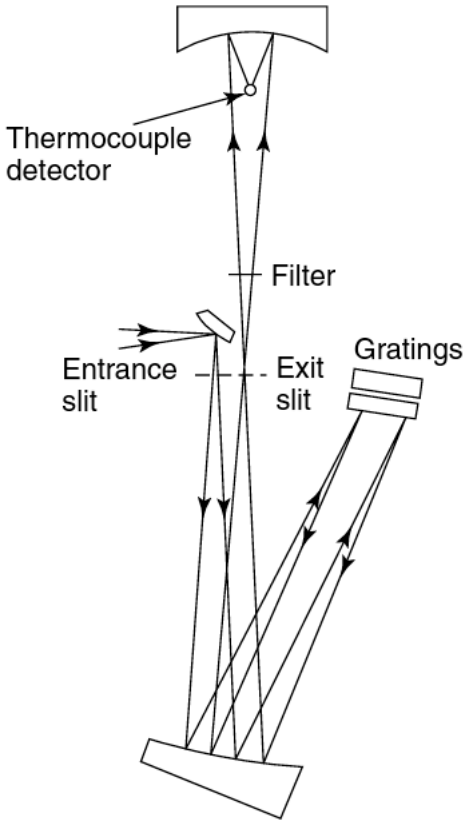


Figure 1.9: Diagram of the optical path of a disperse spectrometer with a grating monochromator. Taken from Stuart, B. 2004 [20]

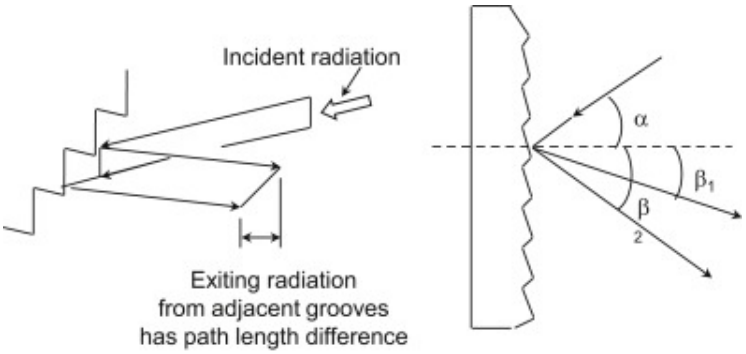


Figure 1.10: Schematic of a diffraction grating. **Left:** Parallel beams from two adjacent grooves are displaced by the path length difference. Constructive interference can occur only when the path length difference is equal to the wavelength multiplied by an integer (first, second, third order). The polychromatic radiation will be diffracted at different angles, resulting in spatial wavelength discrimination. **Right**: (α) is angle of incidence, (β) is the angle of reflectance of the radiation. Adapted from Larkin, P. 2018 [42]

1.2.2 Fourier Transform Spectrometer

Fourier Transform Infrared (FTIR) Spectrometer is a powerful tool for spectral analysis. It consists of a light source, an interferometer, and a detector. The light source of FTIR spectrometers is called Globar. Globar light sources are made of silicon carbide, and they are heated to 1650 °C in order to radiate infrared light with a wavelength range of 2 to 25 µm [43]. The radiation from the globar is directed to a Michelson interferometer (Fig. 1.11). The interferometer contains three main components: a beam splitter, a movable mirror and a fixed mirror perpendicular to each other. The radiation is split at the centre of the beam splitter and half of the radiation is transmitted while the other half is reflected. These transmitted and reflected beams go into the two arms of the interferometer and are reflected, where they are combine and interfere at the beam splitter. Finally, the resulting beam goes to the detector. The movable mirror scans different positions, changing the distance between the two mirrors arms. This will create destructive or constructive interference, which will affect the intensity of the resulting beam. The final result is the intensity of the final beam as a function of the difference in distance of the two mirrors, called interferogram [44]. Examples of interferogram from monochromatic, dichromatic, and broadband radiation are shown in Fig. 1.12 left. The interferogram is then Fourier transformed resulting in the spectrum of the light source (Fig. 1.12 right) Fourier-transformation extracts from the interferogram its different frequencies (in this case wavenumbers) and their corresponded intensity [45] by the following equations:

$$S_(\nu) = \int_{-\infty}^{\infty} I(x) e^{+i2\pi\nu x}\, dx = \mathbf{F}^{-1}[I(x)] \tag{1.9}$$

$$I(x) = \int_{-\infty}^{\infty} S(\nu) e^{-i2\pi\nu x}\, d\nu = \mathbf{F}[S(\nu)] \tag{1.10}$$

where S is the spectrum, I is the interferogram, frequency is ν, x is the difference in distance between the two arms of the interferometer. Equation 1.9 converts the interferogram into spectrum. Equation 1.10 shows intensity variation as a function of wavenumber [46, 47].

FTIR spectrometer produces a sample's spectrum by first producing an interferogram without a sample and another interferogram with the sample. The ratio between the two spectra will correspond to the final sample spectrum.

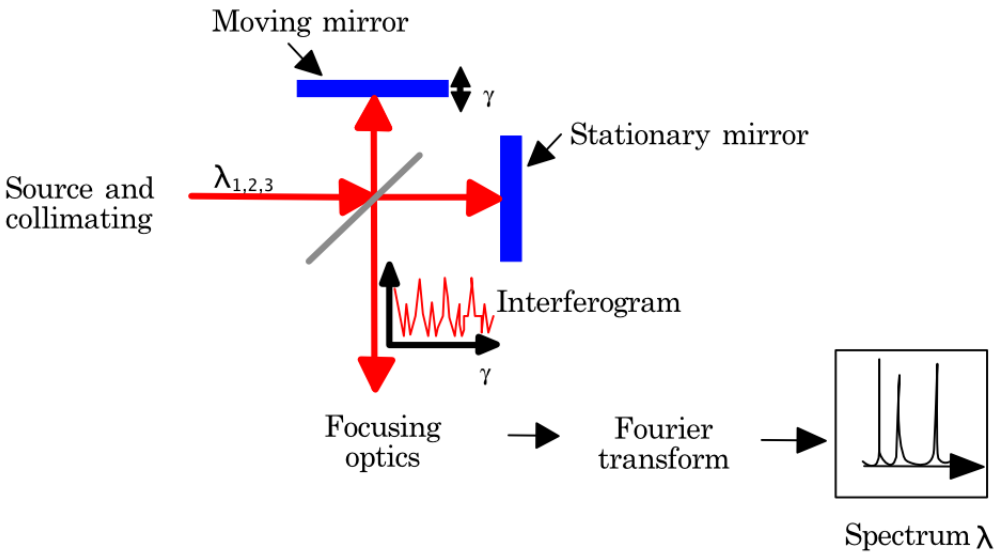


Figure 1.11: Diagram of the Fourier Transform spectrometer with a Michelson interferometer. Infrared radiation with different wavelengths $\lambda_{1,2,3..}$ is directed into the interferometer and the resulting light goes into the detector. The moving mirror changes its position (τ). The resulting interferogram is Fourier-transformed into a spectrum. Plot modified from Saptari, V. 2003 [46].

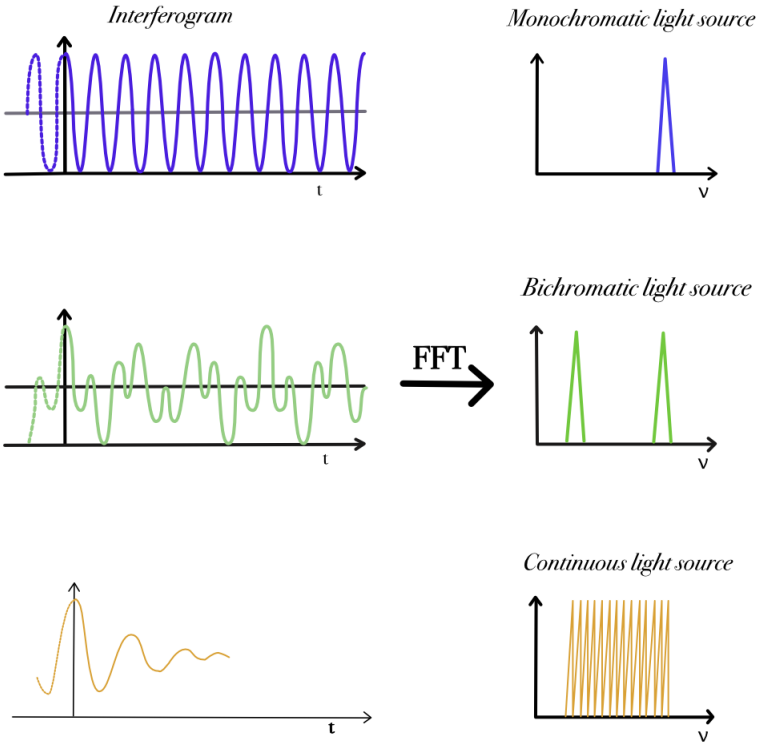


Figure 1.12: Interferograms (left) and its resulting wavelength intensities (right) after a Fourier Transformation for monochromatic (blue), dichromatic (green) and continuous (orange) light sources. Modified from [48]

Advantages

The strengths of FTIR spectrometers are speed, higher signal-to-noise ratio (S/N), and accuracy/reproducibility. Speed is achieved by the ability of the mirror to move short distances very fast, and also the measurement of all wavelengths at the same time [49]. In addition, the use of the Fast Fourier Transform (FFT) which computes Fourier transformations more efficiently, reduces the acquisition time of a spectrum down to milliseconds [47]. High S/N is possible by averaging multiple signals thanks to the rapid scanning of the equipment [20]. Finally, the use of a Helium Neon laser as an internal reference calibration ensures accurate and reproducible measurements [49].

Disadvantages

The disadvantages of FTIR become evident in high spatial resolution image acquisition. Due to the globar light source, there is a lack of spatial coherence and the optical power per diffraction limited spot is low [50]. The low emission intensity of the globar also limits the measurement of samples with a highly absorbing matrix, especially in spectroscopy of proteins in aqueous solutions. The low power limits the path lengths and requires samples with high protein concentrations, which increases the difficulty of using flow cells [51–53]. Moreover, the equipment requires LN_2 detectors, and the cost of equipment can be prohibited to labs with limited resources. Some sampling techniques can require sample preparation (or can be destructive to the sample). Tissues still need long measurement times (in the order of hours or days) when using high-resolution chemical imaging [54].

1.2.3 Infrared spectroscopy sampling techniques

There are several sampling techniques depending on the type of sample to be measured. These techniques can be grouped into transmission methods, internal and external reflectance methods, photoacoustic detection, and gas chromatography/infrared [55]. Each of these methods requires specific sample preparation, which will dictate how easy the technique can be used for a specific project. I am going to focus on the three main ones utilised in this thesis: transmission, diffuse reflectance, and attenuated total reflection.

Transmission mode

Transmission is the most basic method for infrared spectroscopy. This method requires positioning the sample in front of the infrared beam. Here, infrared radiation is passed through a sample. The sample will absorb some of the radiation, and some will pass through or be transmitted [56]. The transmitted radiation is collected with a detector (Fig. 1.13). Transmission can be used to measure infrared absorbance of gases, liquids and thin films [57]. Although liquid and gas samples require no preparation, solid samples need to be prepared. They can be dissolved and examined in solution, prepared as solid suspensions dispersed in a fluid or within a disk of a transparent medium, such as KBr [42]. The absorbance is usually presented as transmittance using equation 1.7.

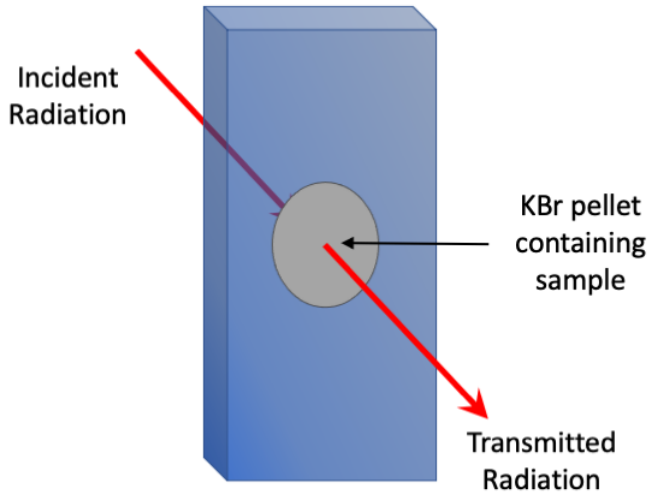


Figure 1.13: Schematic of transmission sampling technique. Sample is positioned in front of infrared radiation. Specific wavelengths will be absorbed by the sample. The transmitted radiation is collected by a detector.

Diffuse Reflectance (DRIFT)

In diffuse reflectance, the incident radiation penetrates the sample, and it is reflected in all directions. Part of the light will be absorbed at sample-specific wavelengths, and the resulting scattering is collected over a large angle. Thus, a diffuse reflection spectrum will contain information about the infrared absorption of the sample. A typical optical configuration for diffuse reflectance spectroscopy consists of a parabolic mirror, which focus the light to the sample, and another set of parabolic mirrors located over the sample which collect the reflected light. The reflected light is then directed towards the detector

[58] (Fig. 1.14). DRIFT is commonly used for powders and fibres [23, 59]. Diffuse reflectance can be expressed as the Kubelka–Munk function (K-M):

$$\frac{(1-R)^2}{2R} = \frac{c}{k} \tag{1.11}$$

where R is the absolute reflectance of the sample, c is the concentration and k is the molar absorption coefficient. However, 1/T, absorbance, ratios, and derivatives has been used to present diffuse reflectance data [37].

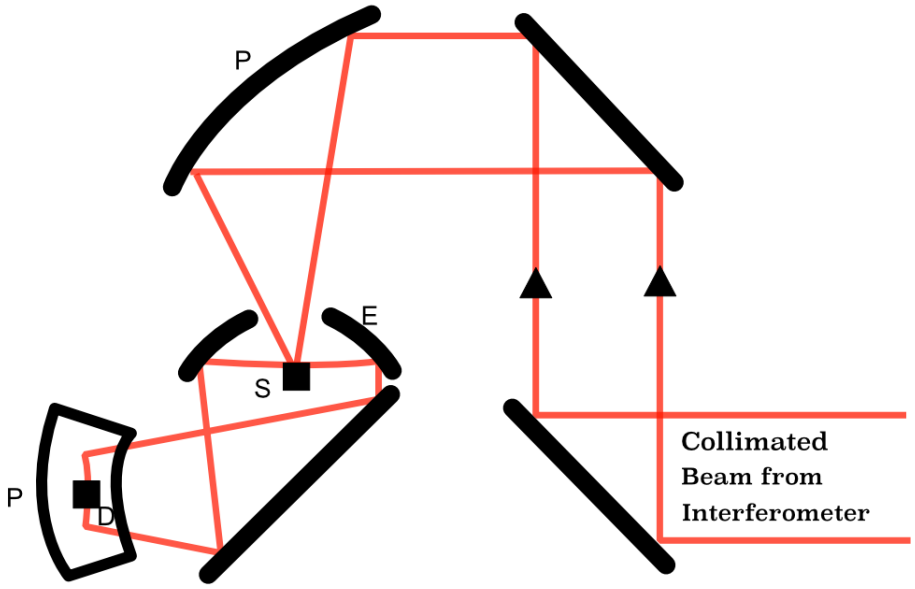


Figure 1.14: Diffuse reflectance optical diagram. Collimated light from the interferometer go towards the sample by parabolic mirror (P). The reflected light from the sample is collected by two mirrors (E) and focused to the detector (D). (Diagram modified from [58])

Attenuated Total Reflectance (ATR)

Attenuated Total Reflectance uses the total internal reflection phenomenon to measure infrared absorbance. It uses a high diffraction, high IR transmission crystal as a medium instead of the air. This crystal offers a surface contact with the sample. The incident beam irradiates the surface trough the optical medium, and the total reflection occurs at the surface (Fig. 1.15). The electric field of the evanescent wave will penetrate the sample, if the sample absorbs infrared radiation at a specific wavelength, the reflectance of that wavelength will be reduced, thus losing energy. if not, the radiation will be reflected completely. ATR measurements are presented as attenuated radiation as a function of wavelength, which will be similar to absorption in transmission mode. This method is

particularly useful for thick samples which are difficult to measure in transmission and requires little to no sample preparation [20, 59].

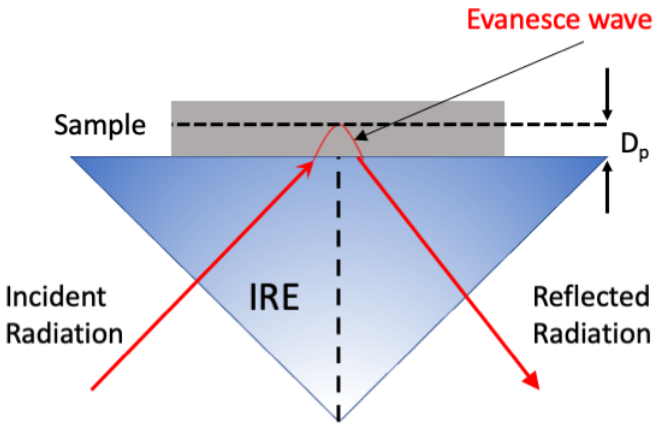


Figure 1.15: Attenuate Total Reflection. The light goes through one of the faces of the Internal Reflection Element (IRE). An evanescence wave generates and enters the sample at a given depth (d_p). The resultant reflected radiation is collected from the other face of the IRE.

1.2.4 Microsampling techniques

One of the advantages of an FTIR spectrometer is that it can be combined with a microscope. This allows us to investigate smaller samples by focusing the light beam onto smaller, localised spots. A FTIR microspectrometer contains a sample x-y stage, two Cassegrain objectives and apertures [60]. The infrared light is passed through a microscope to the sample, and the resulting light is collected by the Cassegrain objective (Fig. 1.16). An aperture is added that focuses the light into the detector. The system also has objectives made of glass for visual inspection of the sample. This hybrid system can be used for transmission, DRIFT and ATR sampling methods [20, 61]. FTIR microspectroscopy can also be used to produce 2D or 3D images of a sample. This is useful for dealing with heterogeneous and complex samples such as cells or tissues [62].

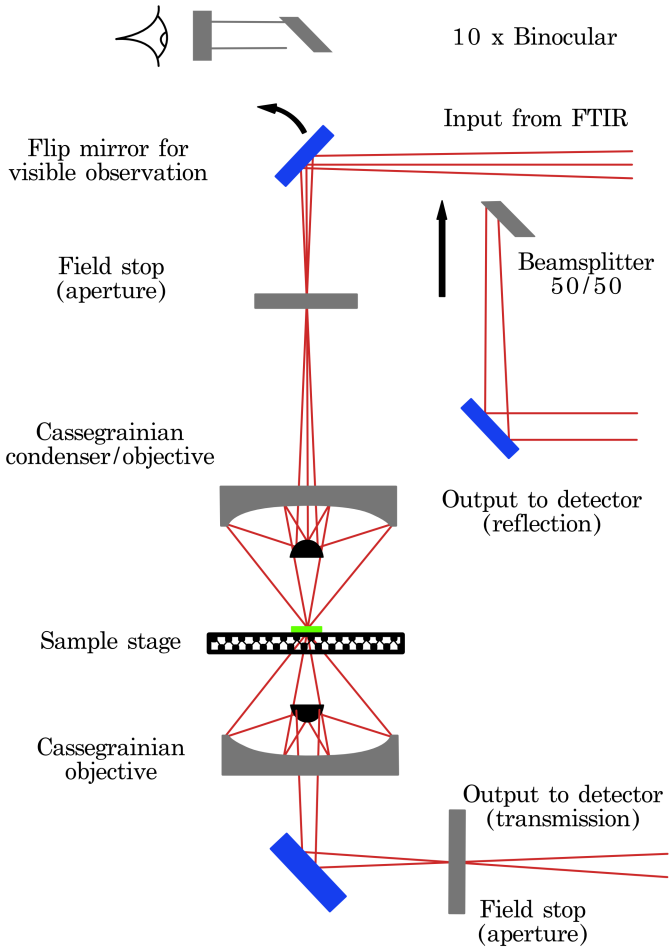


Figure 1.16: Optical diagram of a FTIR micro spectrometer (Modified from Stuart B., 2004 [20] and Katon J., 1996 [60])

1.3 Quantum Cascade Lasers

Quantum cascade Lasers (QCLs) are a type of semiconductor laser developed by J. Faist, F. Caspasso in 1994 [63]. They are based on intersubband transitions and multiple quantum well structures within the conduction band of the semiconductor [64]. Their architecture is based on the theoretical work proposed by Kazarinov and Suris in 1971 [65]. Since their first development, QCLs have seen a steady improvement in design, wall plug efficiency, power, and tuning range [43]; making them a suitable option for mid infrared applications in the fields of communications, security, environmental monitoring, and biology [66–69].

1.3.1 Fundamentals

The structure of a QCL is based on a series of quantum wells composed of thin layers of different semiconductor material with different bands gaps[1](Fig. 1.17a). Inside these quantum wells, one-dimensional confinement occurs allowing band splitting which produces several electronic subbands (indicated as 1, 2, 3 in Fig. 1.17b). The structure can be divided into a gain region and an injection/relaxation region. The gain region will create and maintain a population inversion between the two levels of the laser transition. In general, the active region has three levels, such that electrons are injected in the n = 3 state and the population inversion is maintained between the states n=3 and n=2. The injection/relaxation region follows the gain region, whose role is to transport electrons from level 1 to level 3 of the next period (Fig. 1.17b). Infrared light is generated by an intersubband transition of an electron between the two quantise levels in the conduction band (states n=3 and n=2 in Fig. 1.17b). Then, the electron goes to the next period of the structure by tunneling to the relaxation/injection region (Fig. 1.17). In order to achieve population inversion for lasing, the electrons must be injected rapidly into the upper level 3 and then rapidly extracted from level 2. Fast depopulation of level 2 is achieved by longitudinal-optical (LO) phonon scattering from level 2 into level 1 [70]. Once the electron is in the next period, the process repeats itself across the structure. This process is called cascading [71]. One electron can produce as many photons as periods in the structure [72, 73] which translates to a high optical output power [73]. The energy difference between the two confined energy levels depends on barrier widths of the quantum wells [72], this allows to tailor the emitted wavelength by the layer thickness and material properties [74]. The number of layers ranges from 20 to 35 for lasers in the 4-8 µm, but can reach up to 100 stages [75]. Increasing the number of stages, increase the output power, however, the more periods the devices has also mean larger applied voltages, as the voltage drop per period has to be constant. This requires higher electric power to be dissipated [76].

[1]band gap is the minimum energy required to excite an electron from the valence band to the conduction band

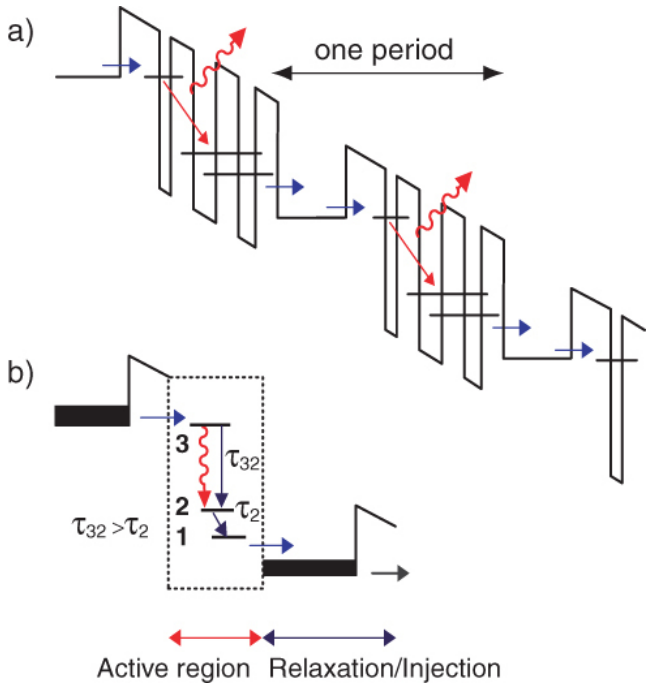


Figure 1.17: a) Schematic of QCL conduction band. Each stage of the structure consists of an active region and a relaxation/injection region. Electrons can emit up to one photon per stage. b) The active region is a three-level system. Photon emission occurs between the 3 → 2 transition. (Diagram taken from Faist, J. 2013 [71])

1.3.2 Fabrication

The fabrication of a QCL consists of epitaxy of the semiconductor structure, waveguide fabrication, device mounting and encapsulation. This process is possible using molecular beam epitaxy and metalorganic chemical vapour deposition [77]. The first step is the growth of the active region, which is made of layers that have high crystalline and chemical purity. The material used for QCLs is chosen to target a specific wavelength. For example, heterostructures for wavelengths between 3 and 24 μm are made of Indium Gallium - arsenide/Aluminium Indium - arsenide/Indium Phosphide (InxGa1-xAs/InyAl1-yAs/InP) heterostructures [78,79]. These materials have higher optical gain, the inherent ability for vertical dielectric waveguiding and the technology to fabricate them is mature [80]. Short wavelengths (≈ 2.6 μm) require higher conduction band offset, which is achieved by using $_{0.53}Ga_{0.47}As/AlAs_{0.56}Sb_{0.44}$ or $InAs/AlAs_{0.16}Sb_{0.84}$. Longer wavelengths, around 20 μm, can be reached suing $In_{0.53}Ga_{0.47}As$ with $GaAs_{0.51}Sb_{0.49}$ barriers. Performance-wise, the best materials are gallium indium arsenide/Aluminium indium arsenide (GaInAs/AlInAs) grow on Indium Phosphide (InP) substrates, Gallium arsenide/Aluminium gallium arsenide (GaAs/AlGaAs) grown on GaAs substrates, Aluminium antimonide/Indium arsenide (AlSb/InAs) grown on InAs, Indium gallium ar-

senide/Aluminium Indium antimonide (InGaAs/AlInAsSb), Indium gallium arsenide/Gallium arsenide antimonide (InGaAs/GaAsSb) or InGaAs/AlInGaAs grown on InP substrates [81].

1.3.3 Quantum Cascade Laser configurations

Quantum Cascade Lasers configurations depends on the resonator cavity[2] design. These configurations can be classified into three main types: Fabry-Perot (FP), distributed feedback (DFB) and external cavity (EC). Each of these designs will change the gain of the active region. The gain will be reduced by the losses caused by waveguide losses, mirror losses from the facets and losses caused by using wavelength selection gratings (external diffraction gratings or Bragg gratings) [71].

Fabry-Perot

The Fabry-Perot is the simplest design used in QCLs. It consists of a QCL chip with high reflection coatings on the end facets of the laser ridge. This configuration has multimode emission over a wide spectral range that has been used in applications where absorption bands are broader, such as liquid analysis [26].

Distributed Feedback Quantum Cascade Laser

Distributed Feedback Quantum laser chips have a Bragg grating integrated into the laser waveguide. The Bragg grating increases losses for all the wavelength except the one that the Bragg was designed for (which is defined by the grating period) [64, 82]. The laser can be tuned within a range of 5 cm^{-1} through changing temperature operation or injection. Both methods can achieve tuning rates of $0.1 - 1.2\ cm^{-1}/K$. Arrays of DFB QCLs can be assembled to cover a wider wavelength range [83]. The narrowband single mode emission is suitable for gas analyses [64, 74]

External Cavity Laser

The external cavity design uses an external diffraction grating as wavelength selector. Wavelength selection (Tuning) is achieved by rotating the grating, which changes the

[2] cavity where the laser is amplified

frequency of the optical feedback of the gain chip. This allows the laser to achieve narrow linewidth and broad tunability with tuning ranges of hundreds of wavenumbers [84]. Since its first development in 2001, where the spectral coverage was 35 cm^{-1} [85], EC-QCLs can now reach 556 cm^{-1} with a single chip [86], and even larger tuning ranges (1000 cm^{-1}) [87] through the combination of several gain chips. The most common configurations for EC-QCLs are shown in Fig. 1.18. In the Littrow configuration, the diffraction grating is placed over a rotation stage and tuning selectivity is achieved by changing the angle of the grating. The first order from the diffraction grating is reflected back to the laser, while the zeroth order is coupled out by a mirror. This configuration has a high efficiency, power and easy alignment [88]. On the other hand, the Littman-Metcalf configuration has the diffraction grating in a fixed position. The first order goes into a mirror, which is reflected to the grating and into the laser as optical feedback. Tuning is achieved by changing the angle of the mirror. The advantages of this configuration are a fixed direction of the output and the smaller linewidth due to stronger wavelength selectivity [89]. External cavity QCL can reach sweep rates up to 1 kHz [90] making them suitable for fast spectral acquisition.

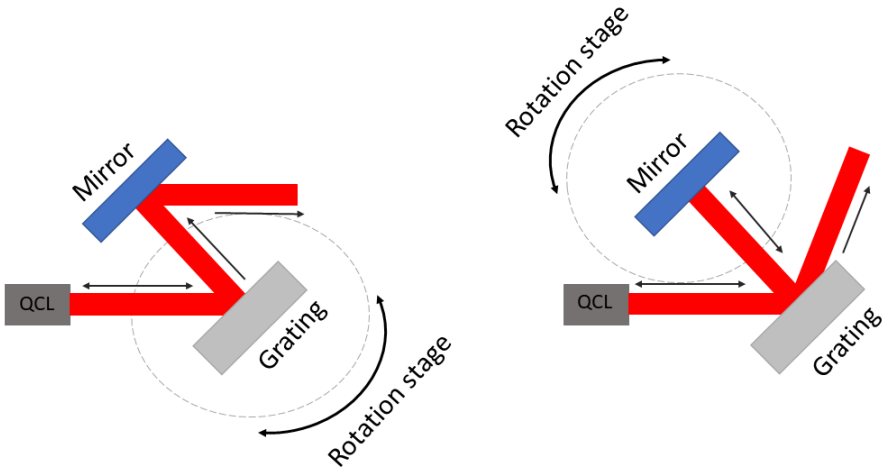


Figure 1.18: External cavity configurations. Left: Littrow configuration. The first order diffracted beam is coupled back into the QCL, the 0th order is coupled out by a mirror. Right. Littman-Metcalf configuration: The first order diffracted beam is reflected off a mirror and diffracted a second time and then coupled back into the laser. The 0th order is coupled out

1.3.4 QCL applications on mid-infrared spectroscopy

The mid infrared region is very important for characterisation of biological samples due to the strong absorption bands from lipids, nucleic acids, proteins, and carbohydrates [24].

Moreover, the presence of high transparency atmospheric windows at 3–5 and 8–12 μm regions are useful in gas sensing/monitoring [91]. The high power, broad tunability and small footprint of QCLs are ideal for spectroscopy applications. Furthermore, the increasing number of companies building these devices have extended the opportunity for testing in a wide area of applications [73]. Here, I will describe some of the main areas of research where QCL are used for mid infrared spectroscopy.

Traced gas detection

Due to global warming over the last decades, greenhouse gas monitoring has become increasingly important, especially for CH_2, CH_4 and N_2O next to CO_2. QCLs with their very narrow linewidth have been widely used for monitoring of such gases. QCLs in Littrow and Littman-Metcalf configurations are the most used for traced gas detection [92]. Detection of N_2O at very low concentration (as low as 2 parts per billion) is achieved using QCLs with multipass cells (gas cells where the laser makes multiple passes, increasing the path length) [93,94] and detection of CH_4 and N_2O simultaneously [95]. Remote gas sensing and stand-off detection of explosives [69,96] and hazardous chemicals [97,98] is also possible using ultra-fast tuning EC-QCLs.

Biomedical applications

Infrared spectroscopy is a promising tool in histopathology, however, the slow speed of data acquisition has limited the use of commercial FTIR into routine application [99,100]. Quantum Cascade Lasers have shown a dramatic decrease in acquisition time compared to FTIR, reducing the processing time from hours to minutes in breast cancer diagnosis [101, 102], colon [103], and esophagus [104] using only key-specific wavelengths and hyperspectral imaging. QCLs have been tested extensively for non-invasive glucose monitoring. Results show systems that can measure glucose concentration from solutions and whole blood using transmission and ATR modes [105–109], directly from the skin (Fig. 1.19) [110, 111] and from the lips [112] with high accuracy. Another biomedical application is breath gas analysis to detect acetone for diabetes diagnosis [113] (i.e., CO and CO_2 [114] and H_2O and N_2O [115]). However, none of these QCL-based systems have been able to transition into routine implementation for clinical use.

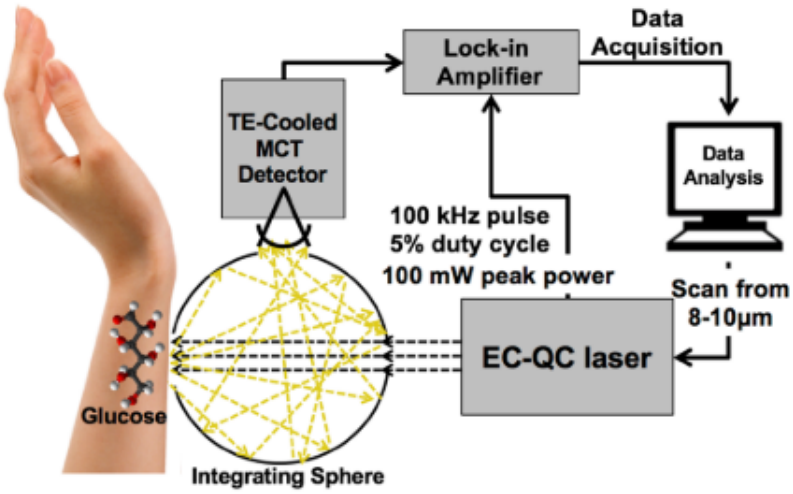


Figure 1.19: Use of QCLs for glucose monitoring by measuring light scattering from skin. Taken from [110]

Polymer identification

Another area of interest for QCLs is the identification of polymers, specifically microplastic characterisation from sea samples. Two important applications have been developed in the last years, the first one using diffuse spectroscopy with QCL in the range of 5.59 - 7.41 μm which can detect five plastic consumer products (polyethylene terephthalate (PET), high density polyethylene (HDPE), low density polyethylene (LDPE), polypropylene (PP), and polystyrene (PS)) with a 97% accuracy [116]. The other application used a commercial chemical imaging microscope (SPERO QT, DRS Daylight Solutions) with four QCL chips covering from 1800 to 950 cm^{21} for fast hyperspectral imaging of environmental samples (e.g., marine surface water, deep sea water and snow [117](Fig. 1.20). They achieved better results in the number of particles per sample identified due to the higher resolution of the QCL system. These studies show how quantum cascade lasers can be incorporate into the process of plastic identification and quantification.

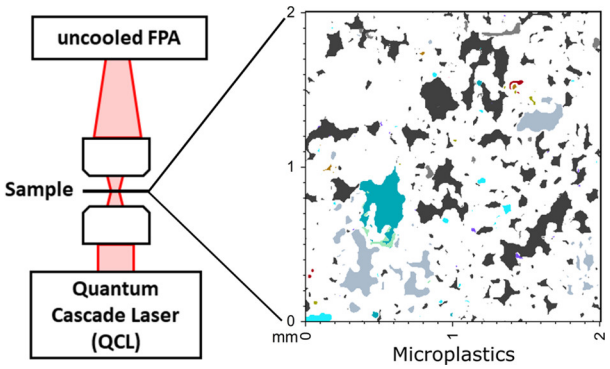


Figure 1.20: Microplastic characterisation of seawater samples using a commercial QCL chemical imaging microscope. Taken from [117]

So far, the value of Quantum Cascade Lasers is the ability to generate spectral data faster and with higher resolution compared to its counterpart FTIR for a variety of applications. While this enables collection of high volumes of spectral data, there is a challenge of finding the best approach to analyse it. This is especially important when the aim of the analysis is diagnosis of diseases, identification of different sample types or predicting the concentration of an analyte. The current approach in chemical analysis is the use of linear models such as Partial Least Squares and Linear Discriminant Analysis. They offer a robust and effective way to analyse high dimensional spectral data [118]. However, less popular non-linear techniques can improve classification or prediction when applied to data [119]. Furthermore, the popularisation of more user-friendly software for machine learning analysis has opened the opportunity to test more advanced algorithms. While my thesis focuses on exploring new technology for spectroscopy, it also deals with the exploration of traditional chemometrics and more advance machine learning algorithms to analysis infrared data. Therefore, I will review the basic concepts in the area of machine learning.

1.4 Machine Learning

Machine learning (ML) is an area of computer science which uses algorithms to make decisions using data with minimal or no human intervention [120]. This is possible by using algorithms that can adapt to presented data and new data. The adaptation is achieved by "experience" which makes them better every time the process is repeated. This process is called training, where inputs of data and outputs plus a set of parameters are feed to the algorithm [121, 122]. The algorithm optimised itself to make decisions, but also has the ability to produce new decisions based on new data [123, 124]. A general

approach of using ML is described in Fig. 1.21. Machine learning has been applied to solve problems in several scientific fields such as biomedical sciences [125], chemistry [126], pattern recognition [127], ecology [128].

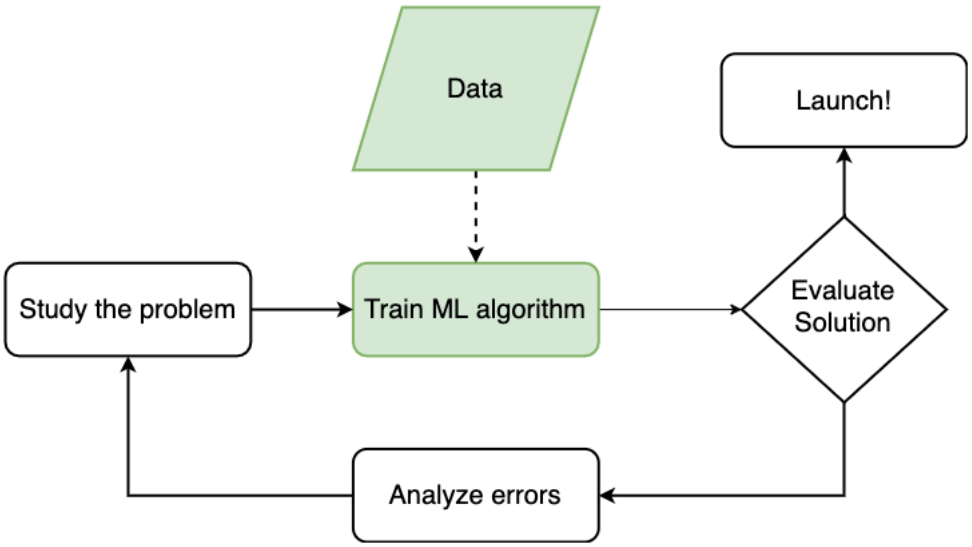


Figure 1.21: Machine Learning approach (modified from Geron, A. [124])

1.4.1 Elements of machine learning

A machine learning project requires a set of elements in order to be implemented. The most important is data, which the core of an ML project. Data can be anything from pixels from images, spectroscopy data, sensor data, measured medical data, and it is denoted as x. In addition to the input data, there is the output value y, which can be the name of the image, a patient's diagnosis or type of sample [120, 129]. The input data are also called 'features', while the outputs are called 'labels'. The relationship between the input and output can be denoted as

$$y = f(x) \tag{1.12}$$

The main objective of ML is to find the function that describes this relationship by using a model. A model provides a framework which allows to approximate the behaviour of a system and make predictions. A model can be expressed by:

$$\hat{y} = \hat{f}(x; w) \tag{1.13}$$

where $\hat{f}$ depends on parameters w to describe the input-output relationship. To find the parameter values, a learning algorithm is used $w = w(x, y)$, also called the loss function (L). The loss function calculates how far (or close) the value of the model is from its

true value. One of the common learning algorithms is the gradient descent, which is an iterative algorithm that helps find the function parameters that minimise the loss function. These calculations help find the parameters of the function in the training data, but the ultimate goal is to predict future data. The model needs to be able to generalise (make predictions with unseen data), therefore, the model is evaluated by assessing how close the predictions are with new data [129].

1.4.2 Types of Learning

Machine learning approaches can be divided into two main areas: Unsupervised learning and supervised learning.

Unsupervised learning

Unsupervised learning does not use label data to train, in fact it does not use training data sets. This approach uses the raw data to find relationships or patterns inside the data that has not been previously postulated [130]. Unsupervised learning can be divided into three problems accordingly to [122]: Association, clustering and dimensionality reduction. Association focuses on discovering the probability of the co-occurrence of items in a collection. Clustering groups items such that items within the same cluster share similar properties compared to items from another cluster. Dimensionality reduction reduces the number of features when dealing with high dimensional data. Dimensionality reduction can be achieved through feature selection and feature extraction [118]. Spectroscopy heavily uses dimensionality reduction for visualisation, cluster analysis or as a pre-processing step before classification. The most common one is multivariate dimension reduction using principal component analysis (PCA) [118]

Supervised learning

Supervised learning is the use of specific data sets that contain the input and output parameters, called labelled data. This labelled data tells the algorithm which results are correct or expected according to the set of inputs. After the training, the algorithm is validated with a test or validation set. These tests or validation sets are not included in the training. If the results are not sufficiently accurate, the process of training and validation are repeated until the algorithm can produce accurate results [131]. Supervised learning

requires the training data sets to be as representative as they could be of the real world data, hence, the algorithm can predict correct results when tested. Supervised learning can be categorised into classification and regression. Classification assigns items into predefined categories. The most basic classification process is binary, where each item can be classified into two categories. Moreover, this classification can be expanded into multiple classification (called multi-label classification). Regression allows the prediction of continuous outputs values from one or multiple predictor variables by fitting a curve to the training data. The most commonly applied supervised learning algorithms are: K-Nearest neighbours, Linear Regression, Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest and Neural Networks [124, 132]. Besides these two categories, there are other types of learning worth exploring such as semi-supervised learning, reinforcement learning, deep learning and batch and online learning [124]

1.5 Mosquito species identification, age grading and insecticide resistance status

Mosquitoes are vectors of numerous important diseases of humans including malaria, dengue, Zika, etc. The ability to measure the potential for, and intensity of transmission of, these diseases depends on surveillance of mosquito populations. The ability to transmit pathogens and susceptibility to control measures varies markedly between mosquito species; including between morphologically cryptic species [133]. Thus, accurate identification of mosquito species is a crucial component of malaria vector surveillance. Vector control programs rely on surveillance of key mosquito traits, among which the identification of mosquito species is crucial [134]. Most of the main vectors of malaria in sub-Saharan Africa belong to the *Anopheles gambiae* s.l. complex and *An. funestus* s.l [135]. There are eight cryptic species (morphologically indistinguishable) in the *An. gambiae* complex [136] with three major vectors: *An. arabiensis, An. coluzzii and An. arabiensis* [137]. While the major vector from *An. funestus* s.l complex is *An. funestus*, other species such as *An. rivulorum, An. parensis, An. vaneedeni* and *An. leesoni* play a role as secondary vectors [138–141]. These species have different biting patterns, blood feeding and resting behaviours which have an impact in their epidemiological importance [142]. Therefore, an accurate species identification is key to apply the most efficient vector control interventions.

Pathogens that are transmitted by insects require a specific time frame to develop inside

their vector before becoming transmissible. The time frame is called extrinsic incubation period (EIP). This period is not fixed across mosquito species, and it will be affected by ambient temperature and humidity [143]. EIP is also used to calculate the vectorial capacity (the ability of a mosquito to transmit a pathogen) in conjunction with mosquito survival, human biting rates and ratio of mosquitoes to humans [9]. Evidence has shown that vectorial capacity increases exponentially when mosquito longevity increases, since older mosquitoes have survived long enough to complete the EIP [143, 144]. Therefore, vector control targets mosquito survival, resulting in a change in age structure with a smaller proportion of mosquitoes surviving long enough to transmit. In order to measure transmission intensity and evaluate the impact of vector control interventions, the age structure of the target vector population must be monitored.

Here, I am going to give a brief overview of the most common techniques for species composition, age grading, and insecticide resistance, as well as new techniques which have been developed in the last years.

1.5.1 Traditional methods for species identification, age grading and insecticide resistant

Species identification

Morphological method The most basic and widely used method is based on morphology. Different mosquito species have specific morphological characters which are used for taxonomic identification. The taxonomist observes the adult mosquito using a stereoscope, looking for those morphological features with the guide of taxonomic keys [145]. It is a cheap method since it requires only a stereoscope microscope, however, it is quite slow, the specimen needs to be in good condition, and it requires very specialise expertise [134, 146]. This technique is not suitable when dealing with cryptic species that are morphologically indistinguishable from each other [137]. That is the case for African malaria vectors, where the major groups *An. gambiae* and *An. funestus* can be identified morphologically, but the cryptic species within each group cannot be identified with this method.

Molecular methods Molecular techniques such as polymerase chain reaction (PCR) provide an alternative for identifying otherwise cryptic species through analysis of species-specific DNA sequences [147, 148]. This technique is widely used and can be applied to

a relatively large number of samples (e.g. each PCR run with a duration of ≈ 30 minutes, and it can process up to 96 samples). However, as any molecular technique, it is not cheap. The price can range for $0.55 up to $10 per sample [149, 150] and it requires specialised equipment and reagents that in some countries a cost prohibited and can take months to arrive.

Age grading

External wear Variation in external wear (i.e: loss of scales and wing fray) to identify young and old mosquitoes was suggested by Perry, R. in 1912 [151]. The technique requires to assesses the minimum wear that female mosquitoes have after its first oviposition. After defining the baseline, samples with less wear can be identified [152, 153]. Accuracy of this technique depends on the specie. For example, misclassification of nulliparous females was 0.9% and 0.1% for *Coquillettidia fraseri* and *Coquillettidia metallica*. Nulliparous females could be recognised by the completeness of wing fringing in *Cx. annulirostris*. However, the collection method was important since damaged samples can result in overestimation of their age [154]. Significant differences in wing damage between parous and nulliparous has also been reported in *Ae. albopictus* [155]. The amount and type of wear was specific for each species. For example, abdominal sternites was better in *Coquillettidia*, while wing fray was better in *Mansonia* and *Aedes* [153]. The disadvantages of this technique are its low resolution (it can only classify between nulliparous and parous), it is subjective, and it depends on the observer's experience.

Presence of mites Gillett, J. [156] suggested the presence of hydrachnid mites on nulliparous females mosquitoes. Nulliparity is only indicated if mites are of the *Hygrobatidae* type and are alive. These mites attached to adult mosquitoes when they emerge from pupae and leave at the time of oviposition [157]. The association between mites and parity depends on the mite specie as well as the mosquito specie. In *Anopheles*, the presence of *Limmensiidae* showed moderate reliability in *Anopheles implexus* (Theo.) from the Zika Forest near Entebbe, Uganda. However, *Arreuridae* was a more reliable nulliparous indicator [158]. Moreover, *Arrenurus* mites were found on only nulliparous *Anopheles annularis* and *Anopheles stephensi*, but in *Anopheles culicifacies*, 36.8% of parous mosquitoes were infested with mites [159]. The effectiveness of the method depends on the percentage of infested mosquitoes in a population, which depends on the type of breeding places. Mosquitoes from breeding sites such as tree-holes, water-storage pots and polluted waters, habitats are unlikely to have mites, whereas mosquitoes breeding in marshes and

weedy ponds are more likely to be infested [151]. For example, in *Anopheles arabiensis*, hygrobatid mites have been found in quite high numbers (up to 39 on a single specimen). Female mosquitoes can be classified as nulliparous if the correct mite family is present and alive.

Presence of meconium Meconium is waste from the larval midgut epithelium located in the midgut of an adult mosquito after pupation. The meconium can remain up to 49 hours after adult emergence, after that, it is degraded by hydrolytic enzymes [160]. It is used to detect nulliparous or young males. Meconium has been found in *Anopheles punctipennis* (Say), *Culex pipiens*, *Aedes triseriatus* (Say), *Aedes aegypti* L., *An. stephensi* Liston, *Cx. pipiens quinquefasciatus* Say [161]. Its use as a nulliparous indicator has been seen in *Culex quinquefasciatus, Culiseta morsitans* and *Culiseta melanura* [151]. It requires minimal dissection expertise. Dissections are performed under a microscope in saline solution and presence of opaque material in the midgut noted. However, it provides limited age information [16].

Presence of larval muscle Larval muscle remnants in adults are visible as a translucent tissue located between the gut and the abdominal wall. They disappear within 12–59 h depending on the temperature (higher autolysis rate at higher temperature). It requires minimal dissection expertise. It provides limited age information [16, 151, 162].

Growth lines on apodemes Evidence of using growth lines on apodemes (inward protrusions of the exoskeleton) for age grading has been used in *Anopheles*, *Aedes* and *Culex* mosquitoes. The mosquitoes have to be macerated in potassium hydroxide (KOH), dissected and then the thorax is cut transversely between the second and third pairs of coxae and up to the scutellum. After that, it stained with haematoxylin. The number of lines observed per apodeme indicates calendar age in days [163]. It requires expertise in dissection and staining. Environmental conditions can limit its applicability [151].

Ovarian tracheation and Polovodova ovariole separation The simplest one is the ovary tracheation method. It identifies whether female mosquitoes have laid eggs (parous) or have not previously laid eggs (nulliparous) by assessing the presence of coiled ends in the tracheoles (skeins) [164]. In general, nulliparous mosquitoes are considered young or unlikely infectious (3-4 days) [165] since they have not taken their first blood meal yet compared to parous [160]. A more detailed approach was developed by Polovodova et

al. [166] to classify mosquitoes by the number of gonotrophic cycles[3] they have completed; providing greater resolution of survival in the parous group. This is based on the observation that each egg oviposition will leave a small dilatation in the ovarioles. Therefore, the number of dilatations on the ovarioles indicates the number of gonotrophic cycles a female mosquito has completed [168, 169]. Morphological techniques are the traditional methods for age grading mosquitoes. These methods are cheap, and they do not require expensive equipment or reagents to carry out. A microscope and basic dissection tools are enough to perform both methods, but these approaches have many limitations and disadvantages. The disadvantages are the low resolution of the ovary tracheation method, the difficulty of counting the number of dilatations, which increases as the more cycles the mosquito goes through [160], and the variability of the gonotrophic cycle length according to the temperature [170–172] which makes it difficult to compare to chronological age. Despise these limitations, these remain the primary methods used in the field.

Ovarian oil injection This method assess the number of dilatations at the base of ovarioles. The process starts with the ovaries removed and then injected with paraffin oil to separate ovarioles. The number of dilatations per ovariole match the number of gonotrophic cycles. It requires a high level of expertise, and it suffers from the same limitations as the Polovodova technique [173].

Parasitic infections This method estimates the survival rate of mosquitoes based on immediate and delayed sporozoite. It required dissecting the mosquito with a needle and further observation with a microscope. Ovaries at Cristophers' stages III, IV and V have increased optical density due to granulation in the basal body which increases with each oviposition and can be used to estimate parity status [174, 175]

Insecticide resistant

Insecticide resistance in mosquitoes gathers several mechanisms including increased metabolic detoxification, which includes the over expression of detoxification enzymes, target site insensitivity where amino acid substitution occurs in the target site of DDT and pyrethroids, and cuticular resistance where changes in the cuticle reduces the penetration of insecti-

[3] A gonotrophic cycle is the process from taken a blood meal, digested it and laid a batch of eggs (oviposition) [167]

cide [176–181]. Not only the mechanisms of resistance are diverse, but so are the genes that regulate these processes [176].

Molecular techniques Several genes are highly associated with some forms of resistance, (e.g. kdr, rdl, ace-1 for target site mutations) [182, 183] which are used frequently in screening vector populations for resistance. However, metabolic resistance DNA markers remains scarce [184–187]. This creates a challenge for monitoring the multiple mechanisms of resistance.

Bioassays Resistance is often assessed in terms of the phenotypic response to insecticide exposure; rather than mosquito genotype. Currently, the gold standard to identify phenotypic resistance is a set of bioassays recommended by the World Health Organisation [188]. Some of these bioassays require processing times up to 24h to test for susceptibility to insecticides, making them slow for fast surveillance. Therefore, the complex mechanisms involve on insecticide resistance remains a challenge for monitoring and evaluation.

1.5.2 Recently developed methods for species identification, age grading and insecticide resistant

Species identification

Image recognition and wing beat frequency These techniques have been tested mainly on *Aedes* and *Culex* mosquitoes [189–191] and a few using *Anopheles* [192]. The differences between wing beat frequency across species can be used to identify mosquitoes. While the technique works well to identify mosquitoes from different genus, the overlap in wing beat frequencies between closely related species remains a challenge. Moreover, any of these techniques have been tested using *Anopheles* cryptic species. The same holds true when using image recognition. The analysis of mosquito photographs using machine learning can identify mosquitoes across different genus and between species such as *Ae. aegypti* and *Ae. albopictus* but struggles with morphologically similar species (i.e. *Culex erraticus* and *Culex pipiens sl*) [193–195]. Therefore, these novel techniques still cannot overcome the complexity of identifying African malaria vectors.

Spectroscopy methods Lastly, Near-infrared (NIRS) and mid-infrared (MIRS) have shown potential for rapid identification of *An. gambiae*, *An. arabiensis* and *An. coluzzii* [17–19, 196–199]. NIRS and MIRS use infrared light to obtain chemical information from the mosquito. This chemical data is then analysed with multivariate analysis or machine learning to predict species. These techniques will be discussed in more detail later in the chapter.

Age grading

Cuticular hydrocarbons Cuticular hydrocarbons (CHCs) are the principal and more abundant components of insect cuticle [200, 201]. CHCs form complex mixtures. Currently, the following CHCs classes have been found: n-alkanes, unsaturated hydrocarbons, terminally branched monomethyl alkanes and internally branched monomethyl, dimethyl, and trimethyl- alkanes [202]. Age grading using gas chromatography in *An. gambiae* have found not only a decrease of n-C21 and n-C23 with age, but also changes across the whole CHCs profile [201]. Similar results have been seen in other mosquito species (*Cx. quinquefasciatus* [203] and *Ae. Aegypti* [204]. Field validation of CHCs with gas-liquid chromatography in *Ae. Aegypti* shows high prediction for ages up to 15 days old [205]. Gas chromatography requires multiple steps and reagents which causes long processing times and increased costs, however, the changes in CHCs in the cuticle with age has opened the opportunity of using other methods such as spectroscopy.

Transcriptional gene profile The expression of some genes is age-dependent, providing an opportunity to estimate age based on expression profiles. Here, gene expression data is used to create transcriptional profiles of mosquitoes of known age to build calibration models. Then, the model is used to predict the age of mosquitoes using their transcriptional profiles [143]. Genome-wide profiles in *An. gambiae* showed changes in 112 genes and 7 candidate genes for age grading [206]. These genes are related to stress-response, detoxification, chitin metabolism, and they encode for oxidoreductases. Further characterisation of AGAP009551 and AGAP011615 genes showed age-dependent monotonic changes in their transcript levels in field populations [207]. Similar results have been obtained with *Ae. aegypti* with three genes (Ae-15848, Ae-8505, and Ae-4274) can predict chronological age (1, 5, 9, 13, 17, 21 days old) [208]. Joy et al. in 2012 [209] found that Ae-15848 was more consistent out of the three genes and can identify individuals older than 14 days old with 87% accuracy and younger than 5 days old with 94.8% accuracy in laboratory and semi field cage samples. This method requires expensive molecular

biology equipment and needs highly trained personnel. RNA extraction for expression profiling requires specific laboratory infrastructure and facilities for sample preservation, making it highly impractical in laboratories with limited resources.

Protein profile Protein expression changes with age in mosquitoes. It has been reported 9 age-dependent proteins in *An. gambiae* and 19 proteins in *An. setphensi* [210]. MALDI-TOF mass spectrometry has been used for age prediction with accuracies over 72% using the same species previously named [211]. In *Aedes*, four protein candidates have shown the potential as good biomarkers for ageing: Actin depolymerising factor, Eukaryotic initiation factor 5A, insect cuticle protein Q17LN8, and Anterior fat body protein (AFP). All of them showed decreases in abundance during ageing, especially AFP [212]. This method requires high-end mass spectrometers and extensive and technically challenging proteomic assessments to find suitable candidate proteins [16].

Spectroscopy methods NIRS, MIRS and Raman spectroscopy have been studied as a rapid way of predicting calendar age. Recently, studies in *Anopheles* [17, 19, 196–199, 213–219] and *Aedes* [220–222] mosquitoes have shown great potential. NIRS measures the chemical composition using wavelengths from 350 to 2500 nm. MIRS measure the chemical composition in the wavelength range of 2500 and 25,000 nm. These two methods can predict calendar age and physiological age with high accuracy. They require no expertise to collect the data, and sample preparation is minimal or none. Both are high throughput with with minimal reagent costs. Raman spectroscopy has been also tested in *Aedes aegypti*, showing good performance in predicting calendar ages (0, 3, 6, 10, 12, 14, 18 and 22 days old). However, the technique has been applied only to homogenised samples [223]. NIRS and MIRS for age grading will be discussed in more detail later in the chapter.

1.6 Infrared spectroscopy application for mosquitoes

The absorption bands detected by NIRS are the overtones of the fundamental vibrations, which are broad and not well-defined (Fig. 1.22). This makes changes in spectra of different compounds less pronounced and therefore, less detectable. Contrary, the absorption bands in MIRS are the product of the fundamental vibrations. These bands are well resolved, and they are assigned to specific chemical groups (Fig. 1.23). Here, I am

going to describe the advances in species prediction, age grading and infection detection using NIRS and MIRS with a review of pros and cons of each method in mosquito surveillance.

1.6.1 NIRS/MIRS application for age grading

The idea of detecting biochemical changes in the cuticle associated with age, species, or infection using NIRS light started to be tested in the area of medical entomology approximately 10 years ago. In 2009, infrared spectroscopy, specifically NIRS (14000–4000 cm^{-1}) technology was first explored as an alternative tool for age grading and species identification methods in mosquito vectors [196]. In this study, discrimination accuracy between *An. gambiae* and *An. arabiensis* was 78.6% on wild mosquitoes and for age grading, discrimination between mosquitoes younger and older than 7 days old was achieved with an accuracy of 80%. Further studies on the *An. gambiae s.l.* complex have generated improved species identification accuracy on wild mosquitoes to 90% [197]. Similar results were achieved in laboratory, semi-field reared and wild caught mosquitoes [19,221]. Replication of these high accuracy results with independent test data sets has been unsuccessful. An extensive study by Krajacich et al. showed a lower prediction power with age groups of less than 7 days of age or greater than 7 days of age and wild caught mosquitoes reared to known ages as independent test sets (accuracy ranged from 49% to 69% depending on the data set used in calibration). These low accuracy values are thought to be the result of the variability introduced by the more heterogeneous diet and genetics of wild caught mosquitoes; and the lack of control of physiological variants during the calibration phase of the model [214]. However, model generalisation has been improved by using Artificial Neural Networks (ANN) with autoencoders to assess parity status with accuracies from 68% to 88.3% using independent sets not used in training [218]. Apart from studies on *Anopheles* species, NIRS can predict age two age groups (<7 or ≥ 7) with 94% accuracy and three age groups (>7d, 7–13d > 13d) with 70.5% accuracy using laboratory reared *Ae. albopictus* [220]. When using field samples, NIRS can differentiate between ($<$ 8d $\geq$ 8d) however, models trained with laboratory samples could not predict field samples and vice versa [217]. A detailed summary of NIRS studies for age grading is shown in Table 1.6.4.

MIRS have recently been used to predict age in mosquitoes. MIRS using ATR has been tested on the *Anopheles gambiae* complex in conjunction with machine learning algorithms. Two species, *An. gambiae* and *An. arabiensis* with age groups of 1, 3, 5, 7, 9, 11,

15 days old were tested with a variety of machine learning algorithms [17]. Age prediction accuracy was variable on each age group and species, with an average of 15% to 97% for *An. gambiae* and 10% to 100% for *An. arabiensis*. Moreover, studies in *Ae. aegypti* have show high prediction accuracy for group ages 2 and 10 days old [222].(Table 1.6.4)

1.6.2 NIRS/MIRS application for species identification

An. gambiae, *An. arabiensis* can be identified using NIRS. It has been tested on field samples with an accuracy of 90% for *An. gambiae* and 76% *An. arabiensis*. High accuracy is maintained with samples preserved in RNAlater and other common preservation methods (ethanol, refrigeration, etc. Table 1.6.4). MIRS with ATR have reported high accuracy for species prediction of *Anopheles* species, from 70 to 85% for *An. arabiensis* and *An. gambiae* [17]. *Aedes* species could be differentiated using MIRS and diffuse reflection using the region of 1800 to 600 cm^{-1} with an accuracy of over 90% to discriminate between lab reared samples of *Ae. japonicus*, *Ae. albopictus*, and *Ae. triseriatus* [224] (Table 1.6.4)

1.6.3 NIRS/MIRS application for infection detection

NIRS has also been used on to assess virus/bacterial infections. *Wolbachia* infections in *Ae. aegypti* were identified from uninfected, with 92 and 88.5% accuracy for females and males. Moreover, identification of different *Wolbachia* strains (wMelPop and wMel) was achieved with accuracies of 96.6 and 84.5% [225]. Zika infection in *Ae. aegypti* were predicted with accuracy of 94.2 to 99.3% [149]. High throughput detection of Wolbachia, Zika and chikungunya in trapped dead female *Ae. aegypti*. Overall accuracies of 93.2, 97 and 90.3% for Zika, chikungunya and Wolbachia respectively [226]. Detection of *Trypanosoma cruzi* (Chagas diseases) in *Triatoma infestans* has also been reported [227].

The ATR-FTIR sampling method has been used in *Ae. aegypti* to diagnose *Wolbachia* infections and for sex and age discrimination. The technique can differentiate between two strains of *Wolbachia*, discriminate between males and females and predict two and ten days old mosquitoes with an accuracy of 90% in lab reared samples. Although promising, accuracy dropped when used to identify infectious individuals in field populations to 78% [222]. Additionally, malaria detection in blood samples in glass slides [228], dried blood in filter paper [229] and blood meal identification in *Anopheles* [230] have been reported.

1.6.4 Advantages and disadvantages of NIRS and MIRS

NIRS and MIRS are trying to complement or replace techniques in mosquito surveillance which are time-consuming, expensive and sometimes inaccurate. The overall advantage of optical spectroscopy methods relies on the speed of sample processing and data acquisition. Sample processing involved minimum or not reagent use, except for desiccant to dry the sample in case of MIRS, this makes the price per sample far cheaper compared to qPCR (110 times cheaper [149]). Data acquisition is fast with an average of 2-3 minutes per sample, making it possible to process 3000 samples per week with current technology [222]. This also decreases price per sample due to the reduction in hours required to process a batch of samples ($10 per sample for qPCR vs $0.10 for NIRS [149]). The data can be analysed with any open-source software such as Python, Orange and R or their commercial counterparts (MATLAB, Unscrambler X). All of this can be achieved with minimum personnel and training [149,197] which makes it an attractive competitor against current molecular biology techniques. The advantages of NIRS relies on in the maturity of the technology, its non-destructive nature, high penetration depth, it can be used in tick samples, in-situ sample measurements and overall cheaper accessories. However, bands in the near infrared region are due to the overtones and combination bands (combination bands $\nu_1 + \nu_2$, second order and third order combination bands such as $\nu + 2\nu_2$), therefore, it has many overlapping bands making it difficult for band assignment (Fig. 1.22). On the other hand, MIRS bands are well resolved, and they are assigned to specific chemical groups. This makes spectral assignment easier and allows extracting more information about the sample, such as protein conformation. Even though, ATR is destructive with the sample, it eliminates any sample processing, increase the measurement speed, and it can collect spectra from thick samples. The main disadvantages of MIRS are the inability to automate spectral acquisition when using ATR, the presence of water in the sample can affect the measurements, and the use of liquid nitrogen detectors (Table 1.5). However, the development of better light sources and detectors can overcome some of these problems associated with MIRS. MIRS has a 7-year disadvantage in terms of being tested in as a tool for mosquito surveillance compared to NIRS, but it has already showed promising results which can complement NIRS. There is still room to explore it with different spectral acquisition techniques, sampling other mosquito body parts, identifying other biological traits (insecticide resistance) and the implementation of new light sources. All of these aiming to improve the technique in order to implement it as a routine tool in mosquito surveillance.

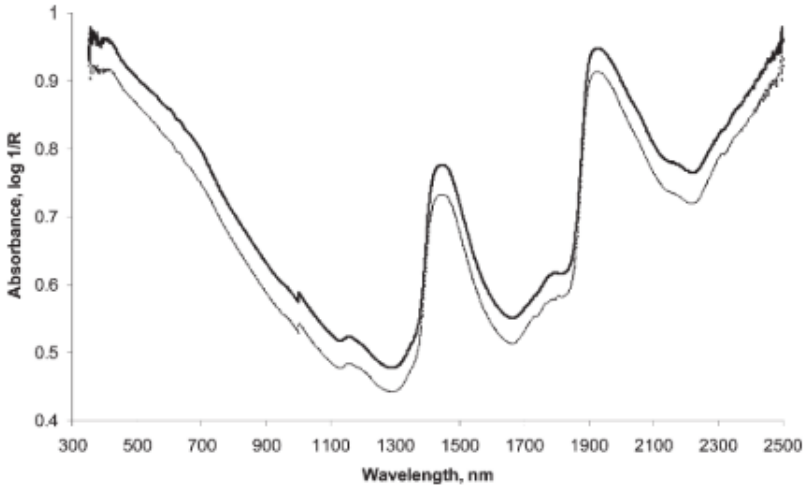


Figure 1.22: Example of near infrared spectra from two species of mosquito *An. gambiae* (bottom) and *An. arabiensis* (top). Broad bands in the NIR region are the result of overtones and combination bands (combination bands $\nu_1 + \nu_2$, second order and third order combination bands such as $\nu + 2\nu_2$) (Plot taken from Mayagaya, V. et al. 2009 [196])

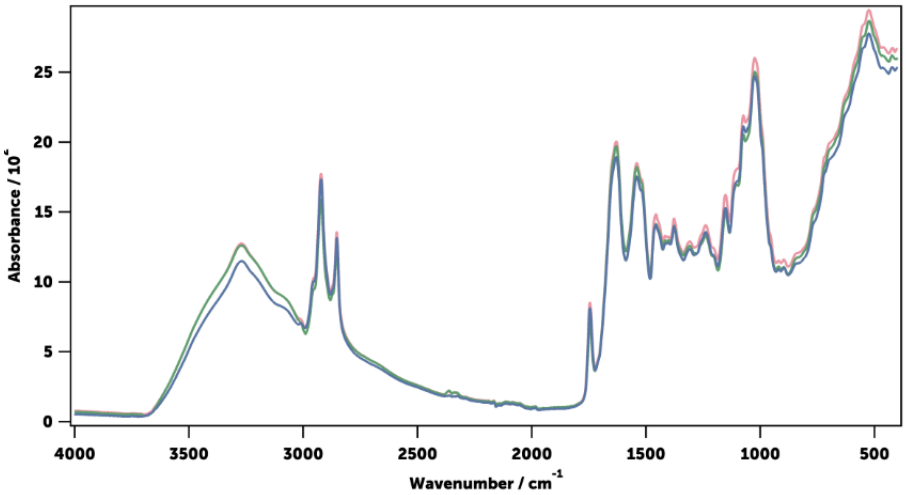


Figure 1.23: Example of mid infrared spectra collected using ATR-FTIR of *An. gambiae* mosquito of 3 days (blue), 6 days (green), and 11 days (pink) after collection. Bands in the mid infrared region are more resolved. This is due by measure fundamental vibrations of molecules (Plot taken from Gonzalez-Jimenez, M. et al. 2019 [17])

Table 1.2: Summary of NIRS studies for species identification. Reported accuracy, chemometrics/machine learning algorithms used and type of sample (laboratory reared, semi-field reared or wild caught samples) are shown. All studies summarised here appear in chronological order

Accuracy	Data Analysis	Ref
76% for *An. arabiensis*† and 90% for *An. gambiae*†	PLS	[196]
High accuracy for species prediction in semi-field and field samples. 89% *An. arabiensis***, 91% *An. gambiae*** and 90% *An. gambiae*†	PLS	[197]
Similar accuracy for species identification in fresh samples compared to RNAlater preserved samples. 82.1% for *An. arabiensis** and 81.3% *An. gambiae** fresh. 76.4% for *An. arabiensis** and 83.4% *An. gambiae** preserved in RNAlater	PLS	[19]
Effect of preservation methods. Accuracy of 91.8% *An. arabiensis** 91.0% *An. gambiae**. > 80% average with different preservation methods	PLS	[18]

* Laboratory reared samples

** Semi-field samples

† Wild caught samples

Table 1.3: Summary of NIRS studies for age grading. Reported accuracy, age groups, chemometrics/machine learning algorithms used and type of sample (laboratory reared, semi-field reared or wild caught samples) are shown. All studies summarised here appear in chronological order

Accuracy	Data Analysis	Ref
Age prediction accuracy of 80% *An. gambiae s.s*† between < 7d, ≥ 7d	PLS	[196]
Accuracy of 89% for *An. arabiensis*** and 78% *An. gambiae s.s*** for <7d or ≥7d. Accuracy of 90% for *An. gambiae s.l.*† validated with by dissections.	PLS	[197]

Table1.3 *(Continued)*

Accuracy	Data Analysis	Ref
Effect of preservation methods for prediction of < 7 d, ≥ 7d in *An. gambiae s.s**. Accuracy of 80% preserved in desiccants, refrigeration, RNAlater. 73% in ethanol and 80.9% in carnoy	PLS	[216]
Effect of preservation methods in accuracy for age grading. Average accuracy of 83.2% in fresh samples and 90% preserve in RNAlater for *An. gambiae* and *An. arabiensis*. Age groups: <7d, ≥d	PLS	[19]
Unable to differentiate between a virgin, pregravid, laid one egg batch or laid two egg batches of *An. arabiensis*	PLS	[215]
No difference in age grading accuracy in samples resistance and susceptible of *Anopheles spp.*. using two age groups, <7 d and >7 d). Accuracy of 79% overall	PLS	[213]
Age grading (<8 d and ≥8 d) and Wolbachia detection in *Ae. aegypti** 91% accuracy for non-infected *Ae. aegypti** and 83% with infected samples with wMel and wMelPop strains	PLS	[221]
Variable accuracy for age grading in *An. gambiae*† using independent test sets. 45% to 67% using <7 or ≥7d age groups	PLS, iPLS, enPLS PLS (MASS), VCPA svmLinear, ORF	[214]
High accuracy for age grading in *Ae. albopictus** 94% for two age group (<7 or ≥7 d) and 70.5% for three age groups (>7d, 7–13d > 13d).	PLS	[220]
Better performance of ANN compared to PLS in regression and binary age prediction. RMSE reduction of 1 day in *An. gambiae** and *An. arabiensis***. Increment of 6% in accuracy when using ANN in binary age prediction (< 7, > 7 d) from 93.6% to 99.4% for *An. gambiae** and 88.7% to 99.0% for *An. arabiensis***	PLS Artificial Neural Networks (ANN)	[231]

Table1.3 *(Continued)*

Accuracy	Data Analysis	Ref
High accuracy predicting Parity status. 97.1±2.2% for *An. arabiensis*† and 93.3±1.2% for *An. gambiae*†	ANN with encoder	[218]
Age prediction in *Ae. albopictus* (< 8d ≥ 8d). High accuracy ((AUC=0.88/AUC=0.97) of lab/field models predicting lab/field samples, respectively. Low accuracy of lab models predicting field samples (AUC = 0.60)	PLS and resample PLS	[217]

* Laboratory reared samples

** Semi-field samples

† Wild caught samples

Table 1.4: Summary of MIRS studies for species identification and age grading. Reported accuracy, chemometrics/machine learning algorithms used and type of sample (laboratory reared, semi-field reared or wild caught samples) are shown. All studies summarised here appear in chronological order

	Accuracy	Data Analysis	Ref
Species	Species prediction in lab reared females was 76.8% in *An. gambiae** and 76.8% in *An. arabiensis**	Logistic Regression	[17]
	Species prediction using the legs and DRIFT was possible with 94.5% for *Ae. japonicus**, 100% for *Ae. albopictus**, 97.9% for *Ae. triseriatus** and 100% for *Ae. aegypti**	PLS-DA	[224]
	Using a few semi-field samples in the training set increased generalisation of CNN model to 90% *An. arabiensis* 100% An. coluzzii and 96% *An. gambiae*	Deep convolutional neural network (CNN) with transfer learning	[199]
	High accuracy age grading in *Ae. aegypti** with group ages 2d and 10d with 0.99 AUC	PLS	[222]
Age	Age prediction of various age groups (1,3,5,7,9,11,15) with accuracies from 15% - 97% *An. gambiae** and 10% - 100% *An. arabiensis**	Logistic Regression	[17]
	Accuracy of 0.98 for 1-4 days old**, 0.91 for 5-10 days old** and 0.96 for 11-17 days**. High accuracy predicting 0, 1 or ≥2 gonotrophic cycles †	Deep convolutional neural network (CNN) and transfer learning	[199]

* Laboratory reared samples

** Semi-field samples

† Wild caught samples

Table 1.5: Summary of NIRS and MIRS principles and pros/cons of field deployment

	How they work	Pros and cons of field deployment
NIRS	Near infrared measures the chemical composition using wavelengths from 350 to 2500 nm. NIR bands are due to overtones and combination modes.	**Pros** Non-destructive to the sample. High accuracy for species prediction and age grading. Miniaturised spectrometers are available. Highly transmitting window to radiation. **Cons** NIR region contains many overlapping bands. Bands assignment is difficult.
MIRS	Mid infrared measure the chemical composition using wavelengths from 2500 and 25,000 nm. MIRS bands are due to discrete fundamental vibrations of biomolecules.	**Pros** More information can be extracted from samples (i.e: protein conformation). High accuracy for age and species prediction. Miniaturised spectrometers are available. New mid-infrared light sources are being developed for field use. **Cons** Destructive when using ATR. Atmospheric absorption limits use in the field. Sample automation cannot be implemented in ATR.

1.7 Aims of the project

In this study, I tested the use of different mid-infrared sampling methods, chemometrics and machine learning algorithms for prediction of key mosquito traits of relevance in malaria vector surveillance: species identification, age grading and cuticular resistance status. I focused on mosquitoes within the *An. gambiae* s.l. complex; as these represent the most important vectors in Africa. Additionally, I assessed the use of quantum cascade lasers as a new tool for spectroscopy in mosquitoes. To achieve this, laboratory and statistical analysis and prototype building were carried out to address the following objectives:

1. Evaluate the use of micro diffuse reflectance spectroscopy for species identification, age grading and cuticular resistance status in *An. gambiae* s.l. mosquitoes. (CHAPTER 2)
2. Asses the use of multivariate statistical methods with spectra collected by Attenuate Total Reflection spectroscopy for species identification and age grading of *An. gambiae*. (CHAPTER 3)
3. Estimate the effect of pre-processing methods and different spectra windows on the accuracy of species identification of female laboratory reared *An gambiae*. (CHAPTER 3)
4. Development of a portable external cavity quantum cascade laser as infrared light source for spectroscopy measurements of mosquitoes. (CHAPTER 4)

This study aimed at improving our understanding of the potential of different spectroscopy sampling methods as well as next generation infrared lasers for mosquito surveillance.

Chapter 2

Diffuse reflectance spectroscopy for predicting age, species and insecticide resistance of *An. gambiae*

2.1 Introduction

Mosquito borne diseases are a major threat for global public health. One of the most important is Malaria change to with an estimated 241 million cases and, 627000 million deaths in 2020. 90% most of which are reported in Sub-Sahara Africa [232]. The parasite is transmitted by mosquitoes of the genus *Anopheles*, with the four primary vector species in Africa being *An. gambiae, An. coluzzii, An. arabiensis* and *An. funestus* [2, 136].

Vector control programs are a key component of malaria management [233–235] and vector surveillance is an important tool to evaluate and monitor the impact of those programs [14, 236]. The constant ongoing pressure from insecticides (mainly pyrethroids) used on insecticide treated nets (ITNs) and Insecticide residual spray (IRS), has given rise to insecticide resistance in most African malaria vector populations [237–239]. This emergence of resistance has the potential to significantly erode the effectiveness of control [240]. Hence, detection and measurement of the degree of this trait in vector populations is also a crucial component of surveillance. Moreover, these interventions have also prompted other ecological changes, including shifts in species composition [4, 241, 242] and behaviour [243, 244]. Several of these changes could also impact the effectiveness of interventions [243, 245], therefore, a constant surveillance of a variety of traits is needed to adapt control strategies to these population changes. [246, 247]

The four main vector species in Africa differ in their behaviour, malaria transmission potential

and response to control, because of that, a good tool for species identification is crucial for malaria vector surveillance [248]. For example, in some areas, *An. arabiensis* is most likely outdoor biting, therefore, its population is less likely to be impacted by the use of ITNs compared to *An. gambiae* s.s [248]. *An. funestus* shows strong resistance to pyrethroids compared to *An. arabiensis* [249]. Therefore, it is very important for vector control programs to know what species are present in any given area and how they are changing with interventions. Species identification can be done morphologically by identifying physical features that are unique to each species [250–252], it can be quite cheap for processing large numbers of samples, and it can be implemented in the field. One of the main challenges of identifying species morphologically is the fact that the main malaria vectors in sub-Saharan Africa occur as a species complex whose members are morphologically identical. Moreover, important features for species identification (legs, wings, scales) may be damaged by collection methods (CDC or BG traps), and that a high degree of expertise is required to identify samples at the species level [134]. Other techniques are used to differentiate cryptic species, including molecular techniques such as PCR and DNA sequencing [148,253] or LAMP [254]. These methods are high-throughput, but the costs of reagents and the requirement of specialised laboratory equipment make them expensive to implement and maintain for programmatic surveillance.

In addition to vector species, information on vector survival and age structures to predict malaria transmission and evaluate the impact of interventions [255,256]. Vector age is epidemiologically important because only mosquitoes that survive long enough for malaria parasites to complete their extrinsic incubation period (EIP; 9 to 12 days for *P. falciparum* [257]) can transmit. Older mosquitoes can go through more gonotrophic cycles, each feeding cycle is an opportunity for a mosquito to become infected and lay eggs [9]. The primary outcome of IRS and ITNs is to increase mosquito mortality, especially among older female mosquitoes. IRS prevent the transmission by killing the mosquito after it feeds, and ITNs inhibit feeding before the mosquito can inoculate a person with sporozoites [258]. Vector control approaches based on killing adult mosquitoes have been prioritised in malaria vector control since the first global malaria elimination program in the 1950s based on the prediction from the Ross-MacDonald model, that transmission is most sensitive to changes in adult survival [259]. A successful intervention will follow with a reduction in the old mosquito population to a younger non-infectious population, changing the age structure of the population [260]. Similarly, the erosion of effectiveness in control methods (e.g., possibly due to emergence of insecticide resistance) could be indicated by a shift in the age structure by an increase in survival [261] affecting malaria transmission [7,262]. For these reasons, it is extremely important for control programs to be able to assess and track changes in the age structure of vector populations, both when evaluating interventions and tracking their long-term impact. This makes surveillance of vector age structure and survival important for assessment of transmission intensity and intervention effectiveness. Traditional methods for assessment of vector age rely on dissecting adult females to assess their reproductive status in terms

of whether they have laid eggs previously (parous) or not (nulliparous). Parity rates can be used to classify females into two broad categories of 'young' and 'old'; based on the assumption that females require 3 – 4 days [160] to complete their first gonotrophic cycle and produce eggs; thus nulliparous females are younger [160]. This technique is difficult to perform, therefore, impractical on larger sample sizes. There are many other techniques that have been investigated for age grading [16] including characterisation of cuticular hydrocarbons [201, 263] transcription gene profiles [207, 208, 264] and mass spectroscopy [211] with all of them showing promising results. However, they rely on laborious techniques, expensive reagents and molecular techniques which are not practical for large scale application, therefore, hard to incorporate into routine surveillance laboratories.

Mosquitoes have developed a variety of mechanisms to resist insecticides. The two main ones are target insensitivity by point mutation (kdr) and metabolic detoxification by over-expression of a family of genes such as P450, esterase and GST [176, 178]. Another mechanism of insecticide resistance is cuticular resistance. The cuticle is modified by an increase of hydrocarbons in the mosquito epicuticle [178]. This is caused by the over-expression of CYP4G16 and CYP4G17 genes [265]. These modifications reduce insecticide uptake by thickening or changing the composition of the leg cloak in the tarsus [179]. It has been identified using electron microscopy and differential quantification of cuticular hydrocarbons and chitin [179–181, 265]. Point mutations can be detected using molecular methods such as PCR but with more complex metabolic resistant mechanisms and in the absence of DNA markers associated with those, a proper diagnosis at a large scale is impossible. Lately, qPCR [266] has been shown to be capable of diagnosing metabolic pyrethroid resistance, but requires expensive reagents to implement as part of routine surveillance in low income countries. However, there is no standard method as PCR to screen populations.

In the last decade, spectroscopy-based techniques have been developed to identify biological traits in mosquitoes. Spectroscopy uses light-matter interactions to generate information that can be used for classification. Infrared spectroscopy measures the quantity of the interaction between infrared radiation and matter. Molecular bonds can absorb specific frequencies of the infrared light. These frequencies will be specific to each functional group inside the molecule, thus providing a fingerprint of the analyte [20, 23, 59]. The most commonly investigated is Near Infrared spectroscopy (NIRS) with a range from 12,500 to 4000 cm^{-1}. Absorption in this region is caused by overtones and combination of fundamental vibrations of CH, NH, OH and SH functional groups [267]. These molecules are present in a vast range of organic and inorganic materials, plant and animal tissues and insects [267, 268]. The advantage of NIRS is the increased penetration deep into the tissue (from 2 to 3 mm up to 8 mm) [267, 269] which allows non-destructive and high throughput measurements [16]. The increased penetration depth can be helpful to detect viruses and parasites since it collects information not only from the cuticle but also from the inside of the mosquito [270, 271]. NIRS has been tested for age and species classification in a va-

riety of malaria [19,196,198,213,214] and dengue vectors [220,225], Chagas vectors [227,272], culicoides [273], agricultural pests [274–276] and *Drosophila* [277]. Moreover, it has been tested in detecting arboviruses and malaria in vectors [149,226,278]. Nevertheless, near infrared spectra peaks are difficult to interpret due to overtone information and combination of vibrations in the region. This makes NIRS spectra difficult to work with for qualitative analysis [267], to extract information about sample composition or feature assignment to specific chemicals [267,268]. The limitation of NIRS in mosquito surveillance is that age grading can be done when broad categories (young and old) but difficulties appear when narrower age groups are needed. However, this issued has been partially resolved by using more advanced machine learning algorithms [231]. Also, field samples are difficult to classify into age groups when laboratory samples are used for calibration [198, 214], however evidence suggest that model generalisation can be achieved by using encoders and neural networks [218].

More recently, mid-infrared spectroscopy (MIRS) has been explored as complementary to NIRS for vector surveillance. MIRS covers a different spectral range (4000 to 400 cm^{-1}), which encompasses the "fingerprint" region of biological samples. The bands in the fingerprint region are the result of fundamental stretching, bending, and rotating vibration of molecules. Moreover, the peaks from MIRS are more resolved compared to NIRS [279]. These peaks appear in characteristic frequencies of the spectra which facilitates its interpretation, band assignment and qualitative analysis [16, 267]. Contrary to NIRS, the penetration depth is reduced due to its shorter wavelengths ($\approx$ 10 µm penetration into the tissue) [280], therefore, spectral information is collected only from the surface of the sample which in the case of mosquito samples is mainly the cuticle (2 - 5 µm) and part of the mosquito interior. This can be an advantage as it only collects information related to the changes occurring in the cuticle. However, the shallow deep of penetration prohibits its use in a non-invasive way [281]. MIRS has been used recently as a tool to identify biological traits in mosquitoes. First, it was used to identify *Wolbachia* infected mosquitoes and age grading in *Aedes aegypti* [222] and to identify *Aedes* species (*Ae. aegypti, Ae. japonicus, Ae. albopictus* and *Ae. triseriatus*) [224]. In *Anopheles* mosquitoes, it has been used to identify blood meals [230], and predict age and species of laboratory populations [17]. In these studies, the most common method for collecting spectra is Attenuated Total Reflection spectroscopy (described in the Introduction section). Although it is a robust technique for acquiring MIRS from tick samples, this approach is destructive (the sample is pressed against the ATR crystal for maximum contact) and thus renders the sample unavailable for further analysis, and cannot be used in other mosquito tissues such as wings and legs because of their small size, limiting the characterisation of different parts that might be informative.

The potential application of MIRS for wide-scale vector surveillance could be improved through use of alternative non-destructive approach for signal acquisition. One possibility is the use of diffuse reflectance; which allows measurement of the flux per wavelength of light reflected in a scattered manner from a sample [37]. This is achieved by focusing the sample beam by an

ellipsoidal or aspherical mirror and collected by another ellipsoidal mirror at 180 or 90 degrees from the incident beam. Diffuse reflectance spectroscopy involves shining a beam of infrared light onto a sample, with part of the light being reflected from the surface, and the other part entering the sample by refraction. The infrared radiation will be absorbed at specific wavelengths, which depend on the sample's composition, and the resulting scattering is collected [20, 282]. A variation of this technique is micro-diffuse reflectance spectroscopy (μDRIFT), in which a microscopy is incorporated into the system to allow more precise direction of where the sample is taken. This feature enables data to be obtained from small parts of a sample by reducing the field of view according to needs. This ability to see the specific sample and to obtain spectra from small areas makes it possible to analyse tissue-specific features that are prohibitive by other sample techniques such as transmission and ATR- FTIR [283]. This could be a particular advantage for aspects of mosquito surveillance where the trait of interest (e.g., type of insecticide resistance) may be restricted to specific body parts or tissues. This technique is well suited to understand tissue specific signals and while it requires a microscope, potentially increase processing time and cost, if proven successful, its implementation will require additional developments to making appropriate to the field.

This study aimed to explore (1) identify the suitability of different mosquito tissues for analysis by diffuse reflectance micro-spectroscopy (μDRIFT) and (2) evaluate the use of the resulting mid-infrared spectra for prediction of species, age, and the presence of cuticular resistance in African malaria vector species.

2.2 Materials and Methods

2.2.1 Mosquito strains and rearing

Experiments were conducted on two morphologically indistinguishable African malaria vector species: *Anopheles gambiae* (Kisumu and Tiassale strains) and *An. coluzzii* (Ngousso strain). The *Anopheles coluzzii* Ngousso strain from Yaoundé, Cameroon [284] is susceptible to all insecticides. The *Anopheles gambiae* Tiassale strain from southern Cote d'Ivoire [285] is resistant to permethrin and deltamethrin, organophosphates, carbamates and organochlorines, and has showed cuticle modifications to reduce insecticide intake [265, 266]. Resistance in this strain is conveyed by a variety of mutations (L1014F kdr and G119S Ace-1) and the up-regulation of P450 [266]. The *Anopheles gambiae* Kisumu strain from Kenya [285] is susceptible to all insecticides. The strains were reared under standard insectary conditions (26±1 °C, 80%±10% humidity, 12 h light:12 h dark cycle) at the University of Glasgow, Scotland, UK. Larvae were fed on Tetramin tropical flakes and Tetra Pond Pellets (Tetra Ltd, UK). Pupae were transferred into cages for adult emergence. Adult mosquitoes were fed ad libitum on 5% glucose. A cohort consisted of pupae

collected and separated in cages for emergence, and then held for either 3 and 10 days post-emergence for age grading/species discrimination and 1, 2, 3 days post-emergence for insecticide resistant experiments prior to scanning. Approximately 20-30 adult females of each age and strain were collected per cohort. Each cohort was reared and collected at different time points. Three cohorts were used for each species and strain except *An. gambiae* Tiassale which needed four cohorts to compensate for the larger sample size of the other strains, since it was the only resistant strain and for some classification problems, groups with roughly the same sample sizes are needed.

2.2.2 Sample processing

The protocol used to process mosquito samples prior to μDRIFT measurements is described in Gonzales-Jimenez, et al. 2018 [286] with minor modifications. Adult females were separated using a manual aspirator, transferred into holding cups and killed by exposure to chloroform for 20 min. Dead mosquitoes were transferred into silica gel desiccant filled tubes and stored at 4 °C. Samples were kept for between 3-5 days to dry them completely prior to scanning.

2.2.3 Diffuse Reflectance Spectroscopy/scanning

Different parts of the mosquito body were analysed. Individual mosquitoes were placed on their side under a gold mirror, which acts as a sample holder. The sample holder is positioned under the microscope in a XY stage of the FTIR. Subsequently, and to improve sample placement, each part of the mosquito was separated from the body before scanning. For each mosquito, the head, the thorax, the abdomen and one hind leg (the femur) were scanned. Spectra were measured using a Nitrogen purged Bruker (Bruker Corporation, USA) vertex70 with a Hyperion 1000 system with a 15x objective and nitrogen liquid-cooled Mercury Cadmium Telluride (MCT) detector with Globar light. The time taken to scan varied from 10 seconds up to 5 minutes depending on the part of the mosquito being scanned at room temperature. Spectra were taken from the Mid-infrared range (600 to 4000 cm^{-1} at 4 cm^{1} resolution). Background measurements were taken every half hour, while assessing the CO_2 region. The average time needed to separate the body part and place it on the gold mirror was 10 seconds. However, body parts of multiple mosquitoes (up to 5 body parts on each measurement) were placed into the gold mirror to save time. The total time needed to separate the body part, place it on the gold mirror, position the XY stage, focus the microscope and spectra collection was approximately 35 seconds for the legs up to 5 minutes and 35 seconds for head, thorax, and abdomen.

2.2.4 Data Analysis

Import of individual files, assembling datasets, pre-processing and analysis of the spectra were performed in Orange [287] and Python (Python 3.6) with in-house developed scripts. Principal Components Analysis was used to assess data clustering.

Species prediction and age grading

Species prediction and age grading were based on analysis of *An. gambiae* and *An. coluzzii* mosquitoes of ages 3 and 10 days old. First, the data set comprised of 3 cohorts. Then, it was split into training set (80%) and hold out set (20%). Baseline performance of the following algorithms was calculated: Logistic Regression (LR), Linear classifiers with stochastic gradient descent learning (SGD), Linear Discriminant Analysis (LDA), Decision Tree Classifier (CART), Random Forest Classifier (RF), Randomised Tree Classifier (ET), Gaussian Naive Bayes (NB) and Support Vector Machine (SVM) using ten-fold stratify cross-validation with the training set. Cross-validation split the training data set into training and test set. The training set is used to train the model, and then the model is tested on the test set. This process is repeated n times. On each round of cross-validation, the data set is randomly split into a training set and a testing set. This process is repeated several times (n-folds) and the average cross-validation accuracy is used as a performance metric. Stratify cross-validation was applied to maintain the same proportion of age groups and species in the training and test set. The algorithm with the highest baseline accuracy was then chosen for further optimisation. Hyperparameter tuning was used to optimised different parameters. The final optimised model was then evaluated with the hold out set (Fig. 2.1).

An alternative approach was used to overcome the limited sample size. Nested cross-validation provides an unbiased estimation of performance parameter scores independently of the sample size [288]. For nested cross-validation, the whole data set was used. Nested cross-validation has two layers, the outer layer and the inner layer. In the outer layer, data was split into 90% to develop the model and 10% for validation. In the internal layer, the remaining 90% of the data was used for optimisation of the model by hyperparameter tuning using a three-fold cross-validation. The optimised model was then validated with 10% of the data, which was split at the beginning. The whole process was repeated 10 times. Each time, a different 10% and 90% of the data set was selected for validation and model development, respectively (Fig. 2.1b). Mean accuracy from the 10 models was reported.

Apart from accuracy as a performance metric, confusion matrices were also reported. A confusion matrix is a table which describes the performance of a model (Fig. 2.2). It shows how many samples of each class have been predicted correctly by the model. The confusion matrix shows four numbers that are used to define the performance of the model. True Positive (TP) when the

observation is predicted positive by the model and is actually positive. True Negative (TN), when the observation is predicted negative by the model and is actually negative. False positive (FP), when an observation is predicted positive by the model and is actually negative and False negative (FN), when the observation is predicted negative and is actually positive. A perfect model will have values of 1 for true positive and true negative, and values of 0 for false positive and false negative. Confusion matrix are shown as a heatmap with a colorbar on the side to improve readability (Fig. 2.2). Values close to 1 will be darker, while values close to 0 will be brighter. Finally, confidence intervals for variable contribution were calculated by bootstrapping (n=100). All the training and evaluation was performed on Python using the scikit-learn library [289]. Hyperparameter configurations can be found in the appendix section A.1 for each model and classification problem.

Additionally, signal pre-processing is known to increase model performance in NIRS [217, 290, 291] and MIRS [222], especially when using reflection methods [224]. Therefore, to assess whether pre-processing increases baseline performance accuracy of different machine learning models for species prediction and age grading, the following pre-processing algorithms were investigated: Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), Normalisation (NORM), and Savitzky-Golay (first and second derivative with 9, 11, 21 windows sizes)

Prediction of cuticular resistance

Analysis of insecticide resistance was based on the comparison between the lines known to be susceptible (Kisumu, Ngousso) and known to be resistant (Tiassale). First, we performed a multi class classification between the three strains using the hold out method approach described before. Then, another classification was performed, with the strains group together into two groups: resistant and susceptible to insecticide. The idea behind this is that spectra for the resistant line will be consistently different from the two susceptible lines. Therefore, the model should be able to group the susceptible lines together. For this classification problem, the hold out set approach and nested cross validation were applied. Performance metrics (accuracy, confusion matrix) and variable contribution were reported as well.

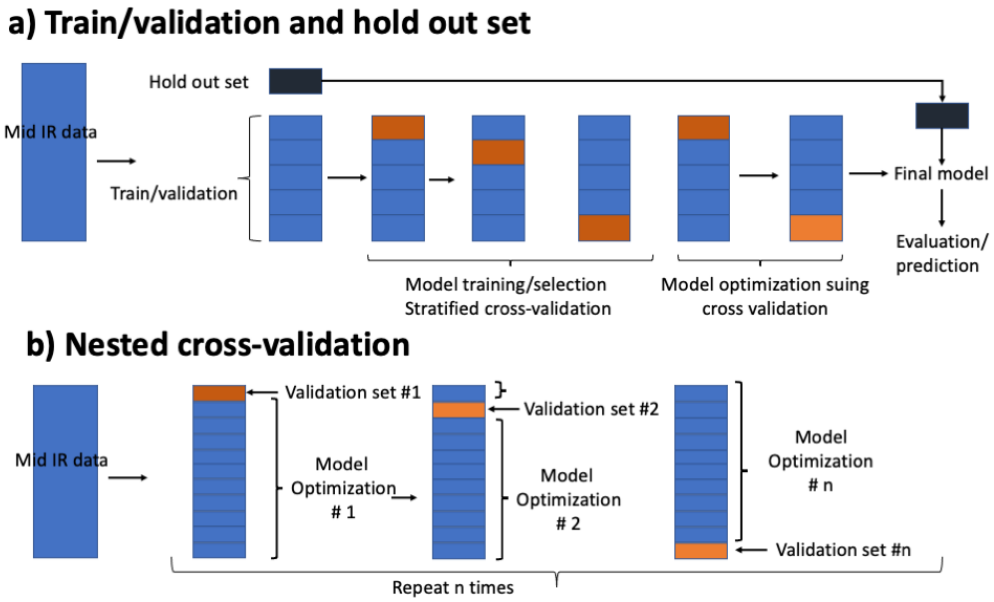


Figure 2.1: Schematic illustration of the process of data splitting for each evaluation method

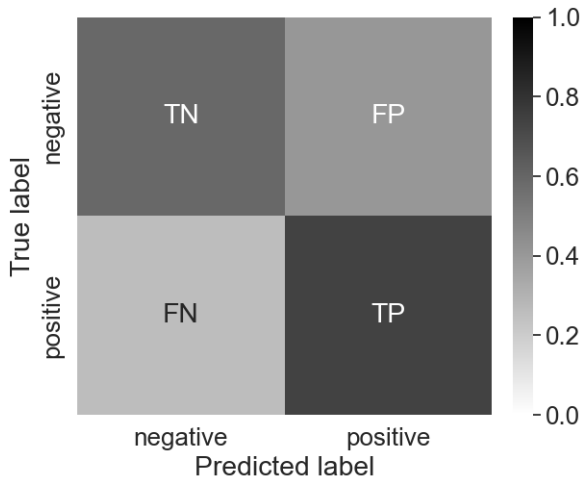


Figure 2.2: Schematic illustration of a confusion matrix as a heatmap. X-axis show the predicted label (positive or negative) from the model. Y-axis shows the actual label (positive or negative). Colorbar on the right side indicates the colormap for each value. TP = True positive, TN = True Negative. FP = False positive. FN = False negative.

2.3 Results

To assess the suitability of different mosquito tissues for diffused reflectance mid infrared spectroscopy, a nitrogen purged Bruker (Bruker Corporation, USA) vertex70 with a Hyperion 1000 system with a 15x objective and nitrogen liquid-cooled Mercury Cadmium Telluride (MCT) detector with Globar light with a spot size of was 50x50 µm. A total of 344 samples from 3 cohorts were used for age grading and species classification, and 291 samples from 3 and 4 cohorts were used for insecticide resistance classification (Table 2.1). Each cohort represents a batch of

mosquitoes of one specie reared to specific ages in one cage. Each cohort was reared, processed and scanned at different time points. Cohorts for species and age grading were reared first. Cohorts for insecticide resistance were reared after species/age grading experiments were completed. The Tiassale strain have 4 cohorts to increase the number of samples to reduced class imbalance when compared to the other strains. All the spectra collected for each cohort were put together in a data set for analysis. The data set was shuffled and split into training set and hold out set. The training set was then used for training and optimisation using cross-validation and hyperparameter tuning. Then, the final optimised model was then evaluated on the hold out set. For nested cross-validation, the whole data set was used and split as indicate in the methods section.

2.3.1 Spectra from different mosquito tissues

Scanning times varied depending on the mosquito body part. Thick tissues (e.g., head, thorax, abdomen) required up to 5 minutes (≈300 scans) of acquisition time to avoid totally absorbed peaks. Even so, absorbance values reached values over 1. On the other hand, legs needed only 16 scans (10 seconds) to obtain a quality spectrum (Appendix A Fig. A.2). Differences in spectra were most notable in the Amide I and Amide II bands (Fig. 2.3 red shaded area) where the peaks of both bands are shifted on thick samples compare to the legs. Spectra obtained from different parts of the mosquito body were notably and consistently different. Specifically, there was a distinct difference between the thicker tissues (head, thorax, abdomen, which clustered together) and the legs (Fig. 2.4). This was also consistent across the different mosquito ages (Fig. 2.5). Subsequent analysis focused on leg samples, given the considerably lower acquisition time.

Table 2.1: Species, strains, and age description of the samples

Species and age classification					
Species	**Strain**	**# samples**	**Age**	**# cohorts**	
An. gambiae	Kisumu	64	3 days old	3	
		105	10 days old	3	
An. coluzzii	Ngousso	69	3 days old	3	
		96	10 days old	3	
Total		**344**			
Insecticide resistance classification					
Species	**Strain**	**# samples**	**Age**	**# Cohorts**	**Group**
An. coluzzii	Ngousso	30	1 day old	3	Susceptible
		30	2 days old	3	
		30	3 days old	3	
An. gambiae	Kisumu	30	1 day old	3	Susceptible
		29	2 days old	3	
		29	3 days old	3	
An. gambiae	Tiassale	37	1 day old	4	Resistant
		39	2 days old	4	
		37	3 days old	4	
Total		**291**			

Note: Each cohort represents a batch of mosquitoes of one specie reared to specific ages in one cage. Each cohort was reared, processed and scanned at different time points. Tiassale strain have 4 cohorts to increase the number of samples and reduced class imbalance when compared to the other strains

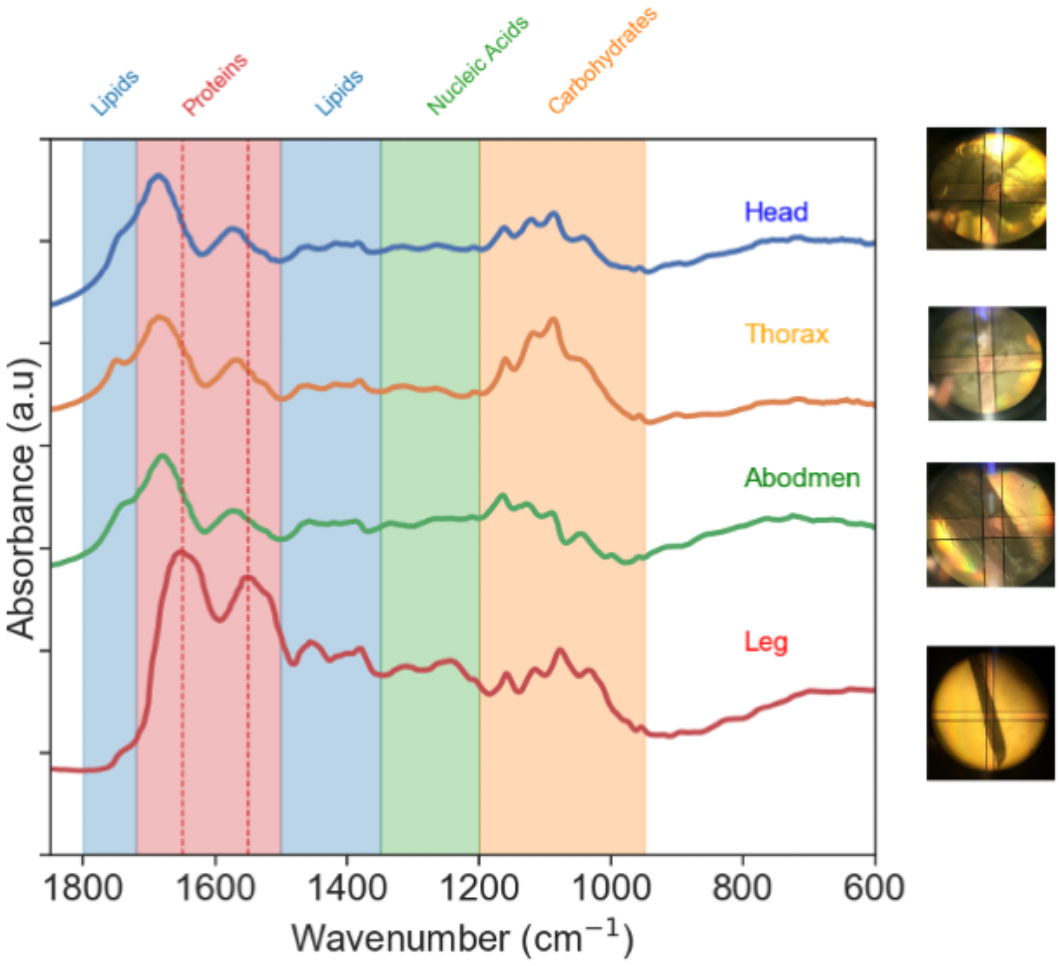


Figure 2.3: IR mean spectra of the head, thorax, abdomen, and leg from mosquito samples. Spectra baselines have been shifted for comparison. Coloured shadow regions show biomolecular peak assignments from 1800 cm^{-1} to 800 cm^{-1} region. Red dashed lines indicate Amide I and Amide II peaks at 1650 and 1550 cm^{-1} respectively. Photos of the field of view from the microscope show the part of the mosquito from the spectra was collected.

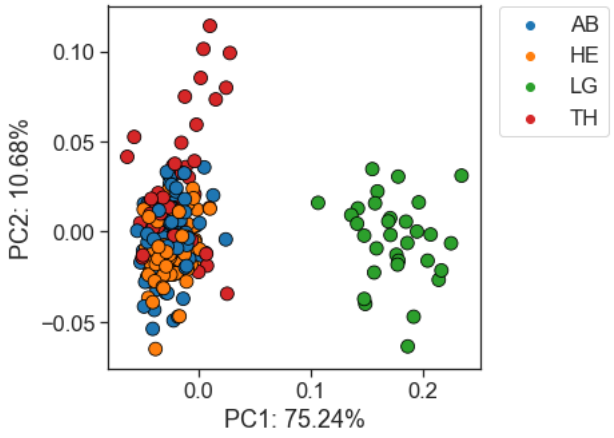


Figure 2.4: PCA scatter plot of spectra data from different mosquito parts using µDRIFT. Legend AB: Abdomen. HE: Head, LG: Leg, TH: Thorax

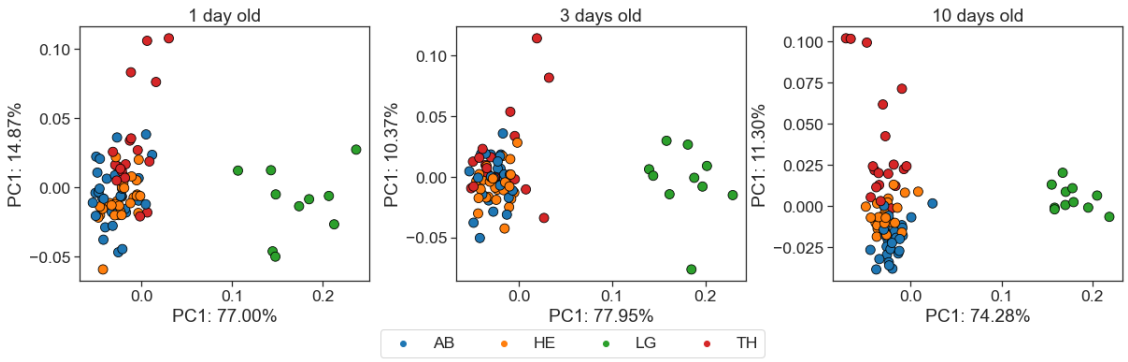


Figure 2.5: PCA scatter plot of spectra data from different mosquito parts using µDRIFT. Legend AB: Abdomen. HE: Head, LG: Leg, TH: Thorax

2.3.2 Identification of *An. coluzzii* and *An. gambiae*

Pre-processing results

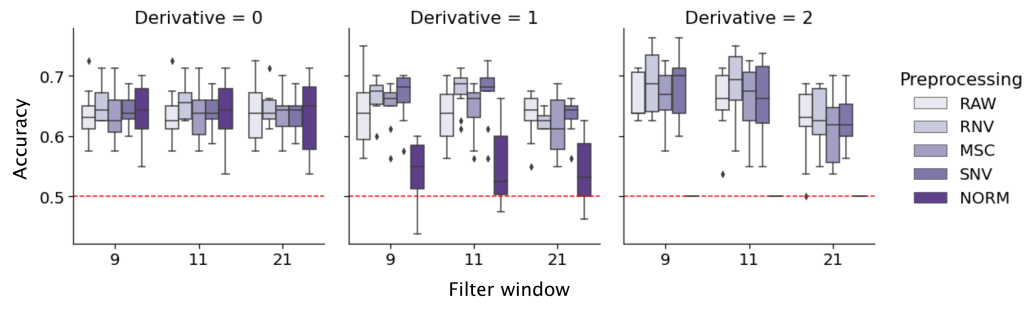
Robust normal variate with Savitzky-Golay second derivative and a 9 window smoothing increased model performance of logistic regression and random forest up to 0.69 and 0.72 respectively (Fig. 2.6a and b). Overall, first and second derivative increased performance of most of the models tested. However, it should be noted that the application of normalisation and Savitzky–Golay second derivative dropped LR accuracy to 50% (accuracy equal to a random classifier). The reason is that extreme pre-processing can erase any difference in the spectra between groups, making it harder for the model to predict species. The same happened with random forest across the different smoothing windows, although not at the same level. Random forest might use the small differences between spectra for classification, but smoothing erases them, decreasing classification accuracy. The differences in model architecture (random forest has a more flexible decision boundary) might help to reduce the effect of extreme pre-processing. Accuracy from the rest of the models can be found in Appendix A Fig. A.5 and Fig. A.6.

Species prediction

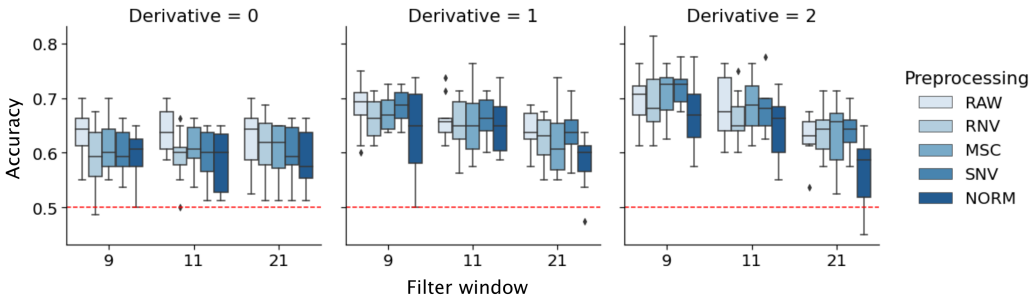
The total number of mosquitoes used in the training set were 231 mosquitoes, with the optimised model being evaluated on a hold out set of 99 samples. Baseline performance of different classifiers with the training data set and ten-fold cross validation is shown in Fig. 2.7a. Logistic regression showed the highest average prediction accuracy of 67.0% ± 5.57% follow by Random Forest. Further optimisation of the model by hyperparameter tuning increased its accuracy to 69.25%. Accuracy using the hold out set was 80.5% (82% for *An. gambiae* and 79% for *An. coluzzii*) (fig.2.7b). Even though, random forest classifier had a higher accuracy compared to logistic regression, when evaluated with the hold out set, accuracy was 73% (Appendix A Fig.

A.4). The top 20 wavenumber values with the highest coefficients are located mainly in the region of 1700 and 1500 cm^{-1}, with some located in the 1300, 1000, 900 and 600 cm^{-1} (Fig. 2.7c). These wavenumber values are assigned to the C=O bond (1700 cm^{-1}) related to proteins and waxes [17], O=C-N bond (1500 cm^{-1}) related to proteins and chitin [292] and C-O bond (1000 cm^{-1}) related to chitin.

Due to the high computational requirements for running nested cross-validation, one model (LR) was chosen based on its high performance obtained from results using hold out set. Model mean accuracy was 69.4% (± 8.8%) after 10-fold nested cross-validation, with a mean accuracy of 69% for *An. gambiae* and 70% for *An. coluzzii* (Fig. 2.8b). This result contrast with the overoptimistic accuracy obtained by hold out set approach. A more balanced accuracy between species was achieved with nested cross-validation compared to hold out set. Model evaluation with hold out set and by nested cross/validation suggests that mid infrared spectra from legs collected with μDRIFT have a signal that can differentiate between these cryptic species; but with relatively low accuracy for *An. coluzzii*.



(a)



(b)

Figure 2.6: Effect of pre-processing in model accuracy Overall species prediction accuracy using the training set and ten-fold cross-validation for **a)** logistic regression and **b)** random forest. Accuracy has been divided into scatter correction algorithm (no pre-processing (RAW), robust normal variate (RNV), standard normal variate (SNV), multiplicative scatter correction (MSC) and normalisation (NORM), filter window (9, 11 and 21 points) and derivative (no derivative, first and second derivative). Red dashed line indicates accuracy of a random classifier (accuracy of 0.5). In boxplots, the horizontal line represents the median, boxes the interquartile range, whiskers the overall range, and black diamonds represent outliers.

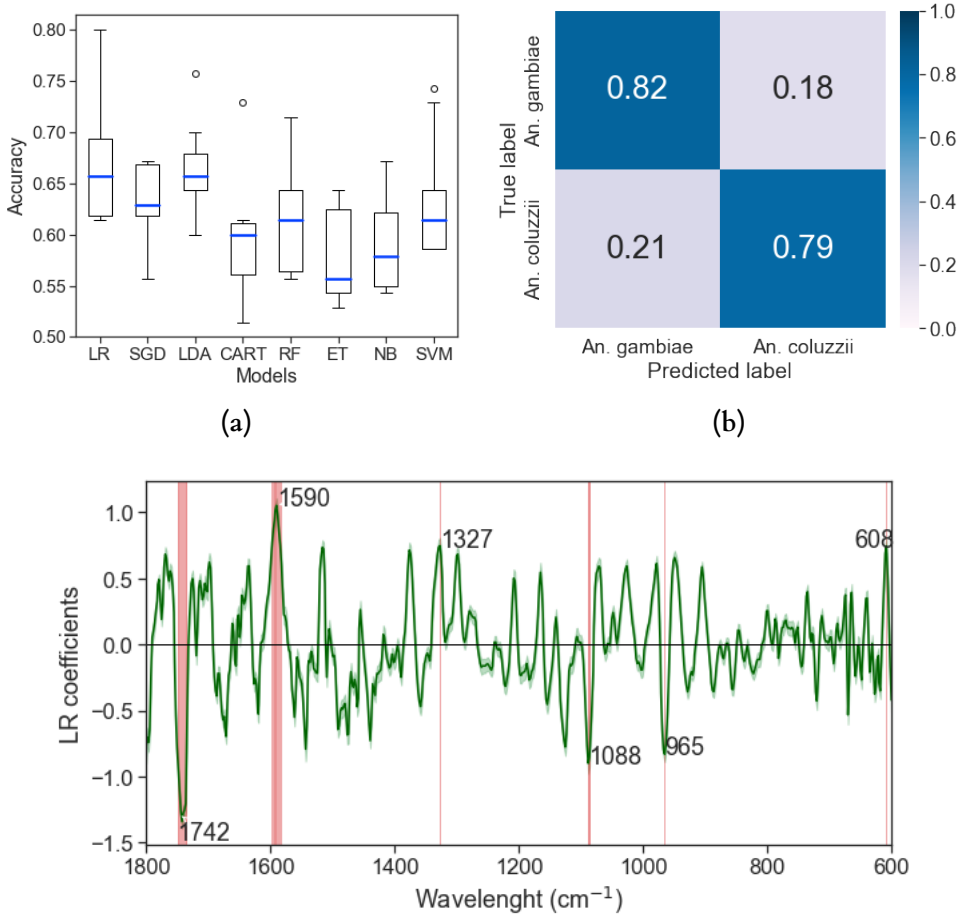
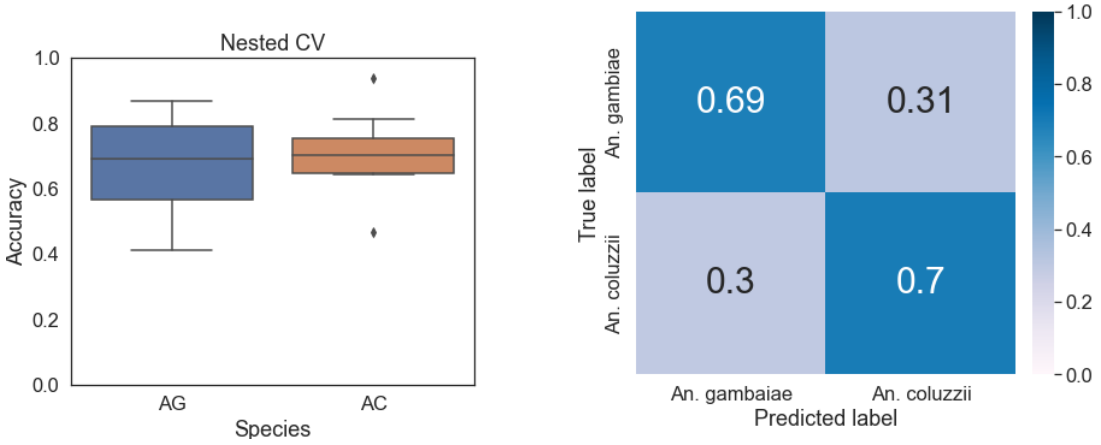


Figure 2.7: a) Baseline performance comparison of common machine learning algorithms. Boxplots show the prediction accuracies for species prediction on training data using different classifiers after ten-fold cross-validation. Models tested included Logistic Regression (LR), Linear classifiers with stochastic gradient descent learning (SGD), Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Random Forest Classifier (RF), Decision Tree Classifier (ET), Gaussian Naive Bayes(NB) and Support Vector Machine(SVM). The best performing model was LR. Boxplot show median (continuous red line) along with the 1st and 3rd quartile, whiskers identify the minimum and maximum values. Circles represent outliers. **b) Confusion matrix of optimised model on hold out set** Normalised confusion matrix of the prediction of the final Logistic regression optimised model using the hold out set. Each row represents instances of true class, while each column represents instances of predicted class. **c) Variable contribution** Variable contribution in the optimised logistic regression model for species prediction. The green line is the average of coefficient values after 100 bootstrap with 95% confidence interval is show in shaded green. Shaded red indicates the location of the top 20 wavenumber values with the highest prediction coefficients. The highest coefficient prediction wavenumber values for each region are annotated



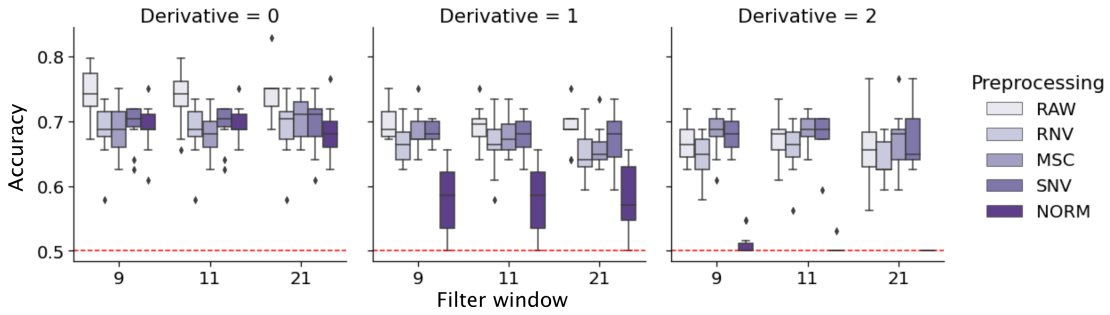
(a) (b)

Figure 2.8: a) Boxplot showing prediction accuracy for each species: *An. gambiae* (AG) and *An. coluzzii* (AC) after ten-fold nested cross-validation. In boxplots, the horizontal line represents the median, boxes the interquartile range, whiskers the overall range, and black dots represent outliers. **b)** Normalised confusion matrix shows the average prediction of ten logistic regression models obtained during nested cross-validation

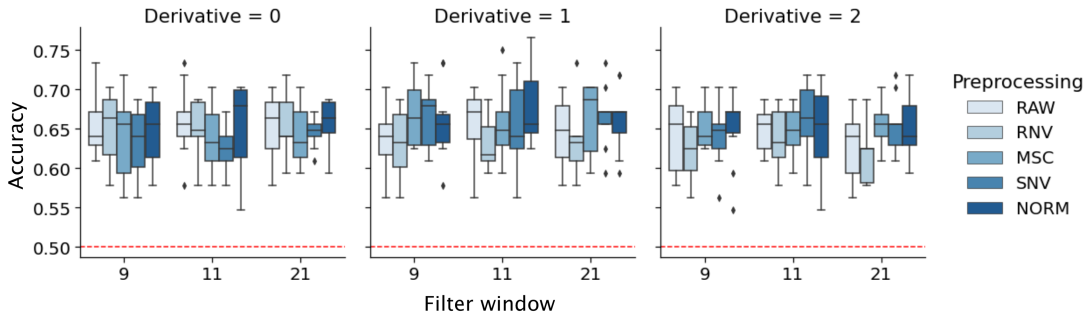
2.3.3 Identification of 3 and 10 days old mosquitoes

Pre-processing results

After assessing species prediction, we used a similar approach for 3 and 10 days age group prediction. We combined both species and split them into the two ages group. The number of mosquitoes used in the training set were 266 mosquitoes, and the optimised model was tested on 53 mosquitoes. Accuracy in cross-validation did not increase with any scatter correction or derivative. A decrease of accuracy started to appear with first and second derivative when applied normalisation. The same tendency of accuracy dropped to 0.5 when using normalisation and the second derivative (window 21 points) was also observed when using LR (Fig. 2.9a) and SGD (Appendix A Fig. A.7). This was not the case with other models (e.g., RF, Fig. 2.9b). The results of all models can be seen in Appendix A, Fig. A.7.



(a)



(b)

Figure 2.9: Overall accuracy using the training set and ten-fold cross-validation for **a)** logistic regression and **b)** random forest for age prediction between mosquitoes of 3 and 10 days old. Accuracy has been divided into scatter correction algorithm (no pre-processing (RAW), robust normal variate (RNV), standard normal variate (SNV), multiplicative scatter correction (MSC) and normalisation (NORM), filter window (9, 11 and 21 points) and derivative (no derivative, first, and second derivative). Red dashed line shows accuracy for a random classifier (accuracy of 0.5). In boxplots, horizontal line represent the median, boxes the interquartile range, whiskers the overall range, and black dots represent outliers.

Results for age prediction

Logistic regression obtained the best baseline performance when predicting 3 and 10 day old age groups, with an accuracy of 73.91% ± 3.13% (Fig. 2.10a). The model was then optimised by hyperparameter tuning, which increased its accuracy to 83%. The accuracy on the hold out set was 87% (89% for three days old and 85% for 10 days old) (Fig. 2.10b). Top 20 prediction coefficients were located in the CH stretch rig stretching (878 cm^{-1}) related to chitin, C=C bond (1485 cm^{-1}) and Amide I region (1680-1690 cm^{-1}), related to proteins and chitin (Fig. 2.10c). This suggests that changes in the cuticle's leg related to age in these two groups (3 and 10) can be detected with μDRIFT and our model can predict with high accuracy for younger mosquitoes. When the model was evaluated using nested cross-validation, there was a decrease in mean accuracy down 77.1% (± 6.5%) after ten-fold nested cross-validation. Class accuracy

ranged from 58% to 100% for 3 days old and 57% to 80% for ten days old (Fig. 2.11a) with a mean accuracy of 82% for 3 days old and 74% for 10 days old (Fig. 2.11b)

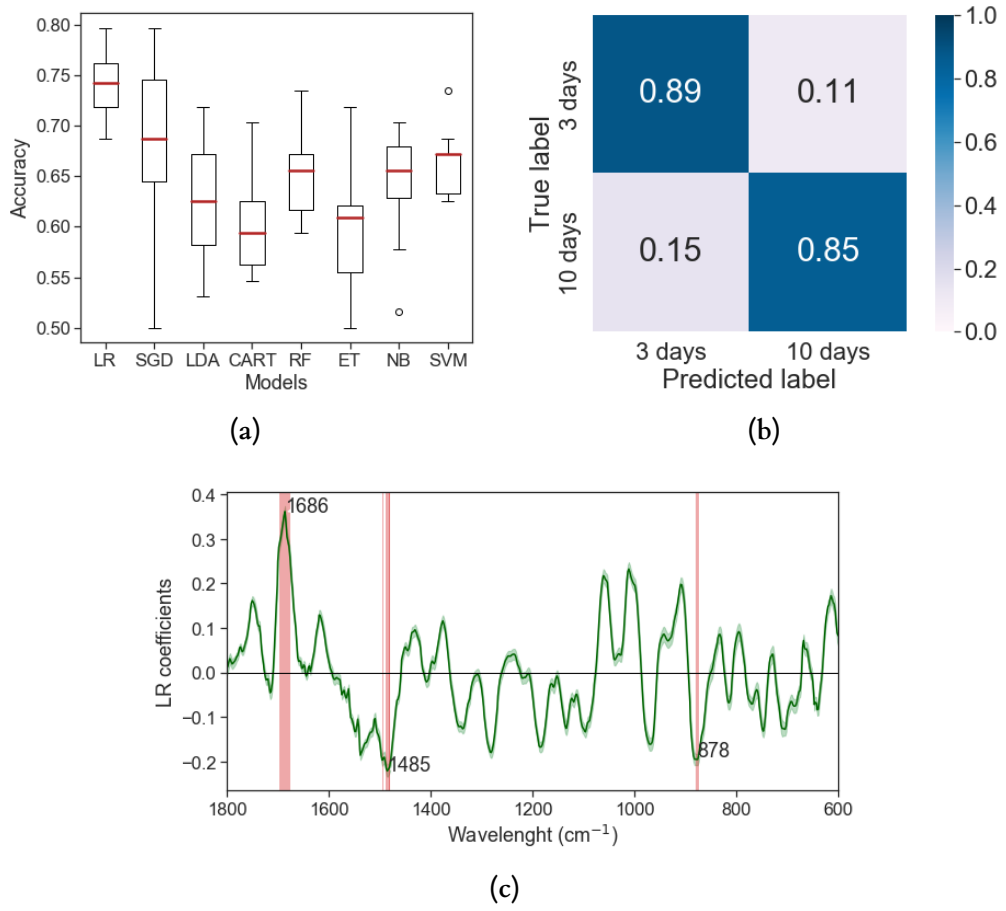


Figure 2.10: a) Baseline performance comparison of common machine learning algorithms. Boxplots show the prediction accuracies for age prediction on training data using different classifiers after ten-fold cross-validation. Models tested included Logistic Regression (LR), Linear classifiers with stochastic gradient descent learning (SGD), Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Random Forest Classifier (RF), Decision Tree Classifier (ET), Gaussian Naive Bayes(NB) and Support Vector Machine(SVM). The best performing model was LR. Boxplot show median (continuous red line) along with the 1st and 3rd quartile, whiskers identify the minimum and maximum values. Circles represent outliers. **b) Confusion matrix of optimised model on hold out set** Normalised confusion matrix of the prediction of the final Logistic regression optimised model using the hold out set. Each row represents instances of true class, while each column represents instances of predicted class.**c) Variable contribution** Variable contribution in the optimised logistic regression model for age prediction. The green line is the average of coefficient values after 100 bootstrap with 95% confidence interval is show in shaded green. Shaded red indicates the location of the top 20 wavenumber values with the highest prediction coefficients. The highest coefficient prediction wavenumber values for each region are annotated

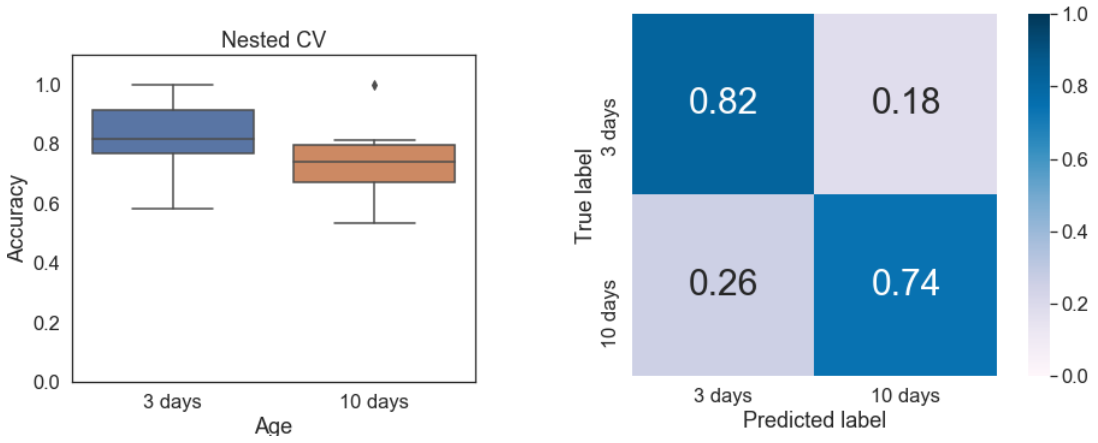


Figure 2.11: a) Boxplot of accuracies for each class: 3 days old and 10 days old after ten-fold nested cross-validation. In boxplots, the horizontal line represents the median, boxes the interquartile range, whiskers the overall range, and black dots represent outliers. **b)** Confusion matrix showing the mean of accuracies of 10 models after nested cross-validation

2.3.4 Identification of cuticular insecticide resistance

After prediction of species and age was performed with supervised algorithms, we tested the same approach to analyses whether μDRIFT could differentiate the strain with cuticular resistance (Tiassale strain) from others (Kisumu and Ngousso strains). A total of 286 mosquitoes were used for this analysis. Initial assessment using PCA analysis showed no clustering between strains (Fig. 2.12a, Appendix A, Fig. A.9). Similarly, there was no clustering when strains were grouped as with cuticular resistant ("resistant") and the other two lines without cuticular resistance ("susceptible")(Fig. 2.12b). The same pattern was observed when comparing each pair of strains to each other (Fig. 2.13).

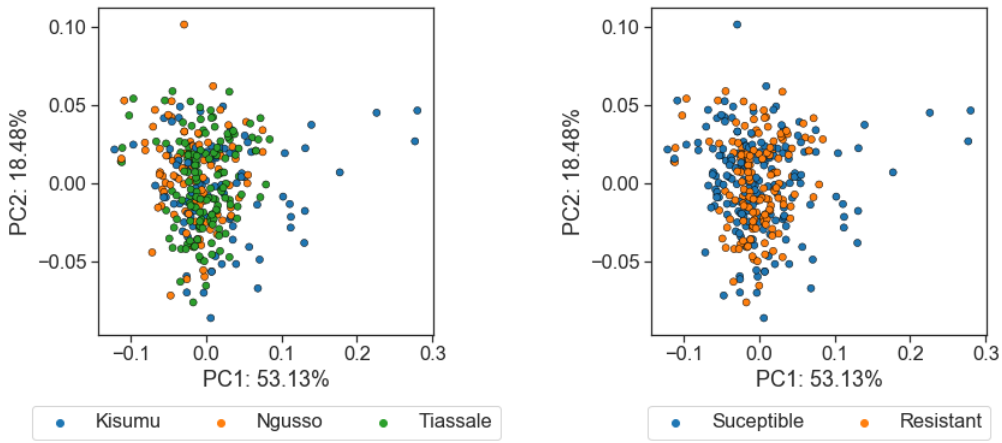


Figure 2.12: a) PCA scatter plot of spectra of each strain (a) and grouped into susceptible and resistant according (b). The variability in the data is explained by the first two PCs in 92.26% of the total

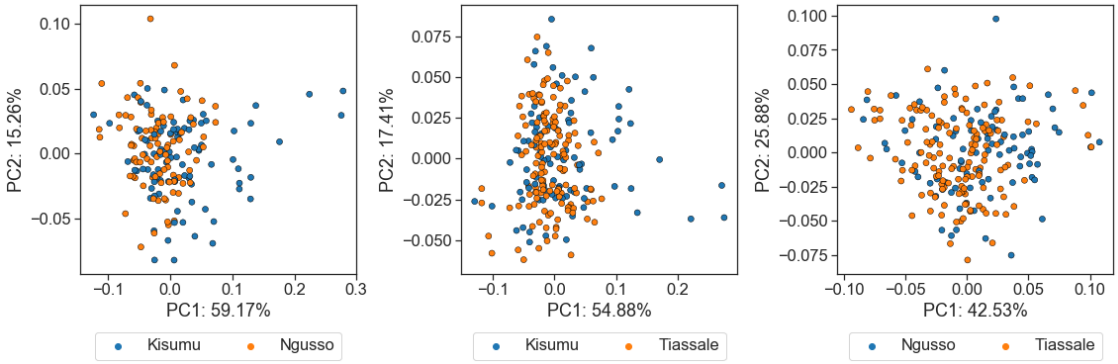


Figure 2.13: PCA scatter plot of spectra for each pair of strains: Kisumu and Ngousso (a), Kisumu and Tiassale (b) and Ngousso and Tiassale

Strain prediction

A multi-class classification for the different strains using SVM was performed. The model used a one vs the rest approach. It involves splitting the multi-class dataset into multiple binary classification problems. A binary classifier is then trained on each binary classification problem and predictions are made using the model that is the most confident. Average accuracy on the hold out set was 60%. The lowest prediction accuracy was for Ngousso with 39% while for Kisumu was 63% and 79% for the Tiassale strain) (Fig. 2.14). Prediction coefficients showed differences in weights for each strain. 1400, 1550, 1650 y 1770 cm^{-1} region for Kisumu strain, compare to 1100, 1300 and 1500 cm^{-1} for Tiassale, and 1200, 1400 and 1500 cm^{-1} for Ngousso (Appendix

A, Fig. A.10). Most of the bands are related to chitin and wax (Appendix A, Table A.1).

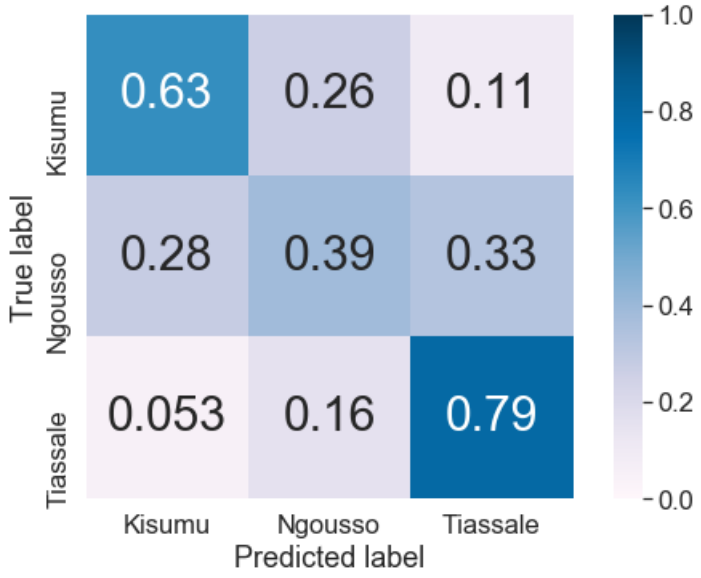


Figure 2.14: Normalized confusion matrices for SVM with "one vs all approach" in validation set on multi-class strain classification. Each row represents instances of true class, while each column represents instances of predicted class

Classification "resistant group" (cuticular resistant strain) and "susceptible" (strains without cuticular resistant)

A total of 200 samples were used for training/cross-validation and 86 samples as hold out set. Mosquitoes were predicted into resistant and susceptible groups with an accuracy of 71.83% ± 4.37% with LR and 73.17% ± 3.20% with SVM (Fig. 2.15a). Although baseline accuracy was higher with SVM, LR showed a higher accuracy in during evaluation with hold out test with 69.76% (72% for resistant and 67% for susceptible) (Fig. 2.15b). Variable contribution of the optimised logistic regression model was assessed using model coefficients (Fig. 2.15c). Top 20 coefficients are located in the C-O stretching (940 – 1130 cm^{-1}) and C-O-C ring (1130-1170 cm^{-1}). Model evaluation using nested cross validation showed a mean accuracy of 71.3% (± 8%) compared to the 69.77% using hold out set. The distribution of accuracy of the model for each class is shown in figure 2.16a. Mean accuracy was 73% and 71% for susceptible and resistant class respectively (Fig. 2.16b)

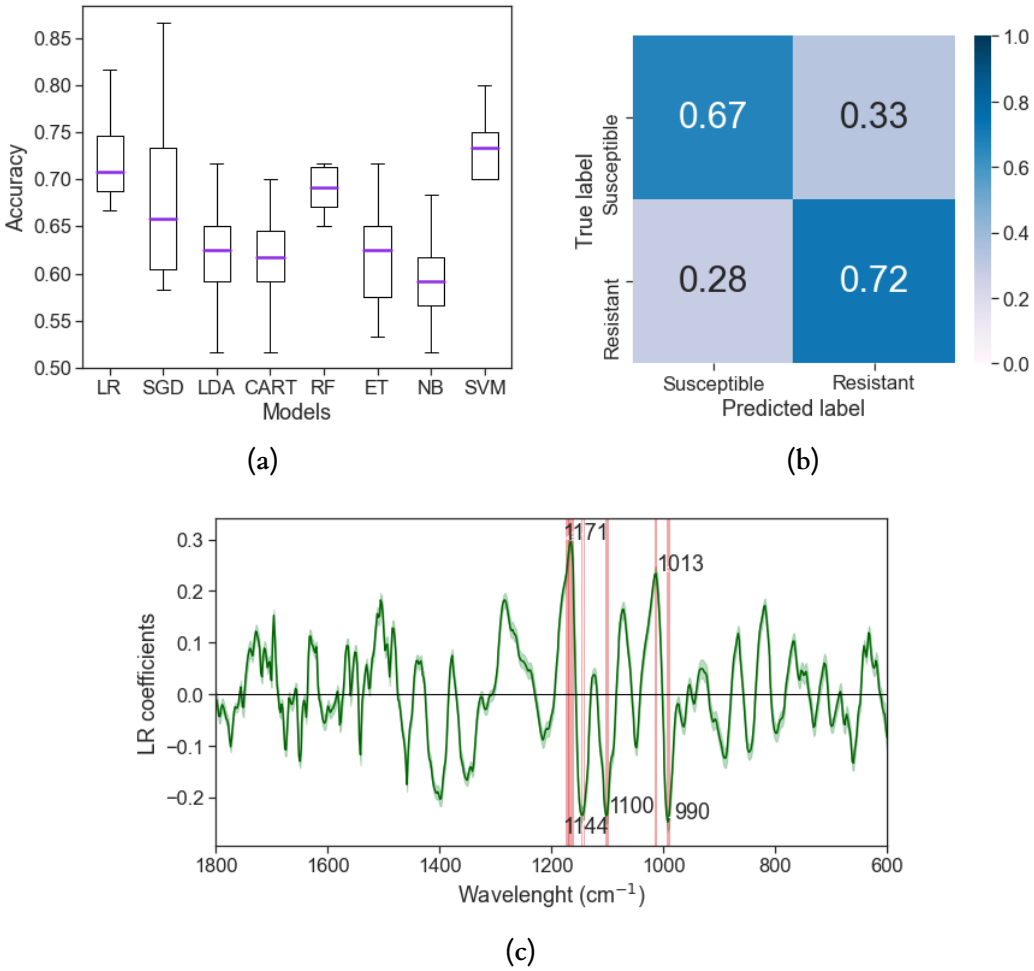


Figure 2.15: a) Baseline performance comparison of common machine learning algorithms. Boxplots show the prediction accuracies for insecticide resistance prediction on training data using different classifiers after ten-fold cross-validation. Models tested included Logistic Regression (LR), Linear classifiers with stochastic gradient descent learning (SGD), Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Random Forest Classifier (RF), Decision Tree Classifier (ET), Gaussian Naive Bayes(NB) and Support Vector Machine(SVM). Boxplot shows the median (continuous purple line) along with the 1st and 3rd quartile, whiskers identify the minimum and maximum values. Circles represent outliers. **b) Confusion matrix of optimised model on hold out set.** Normalised confusion matrix of the prediction of the final Logistic regression optimised model using the hold out set. Each row represents instances of true class, while each column represents instances of predicted class.**c) Variable contribution.** Variable contribution in the optimised logistic regression model for insecticide resistance prediction. The green line is the average of coefficient values after 100 bootstrap with 95% confidence interval is show in shaded green. Shaded red indicates the location of the top 20 wavenumber values with the highest prediction coefficients. The highest coefficient prediction wavenumber values for each region are annotated

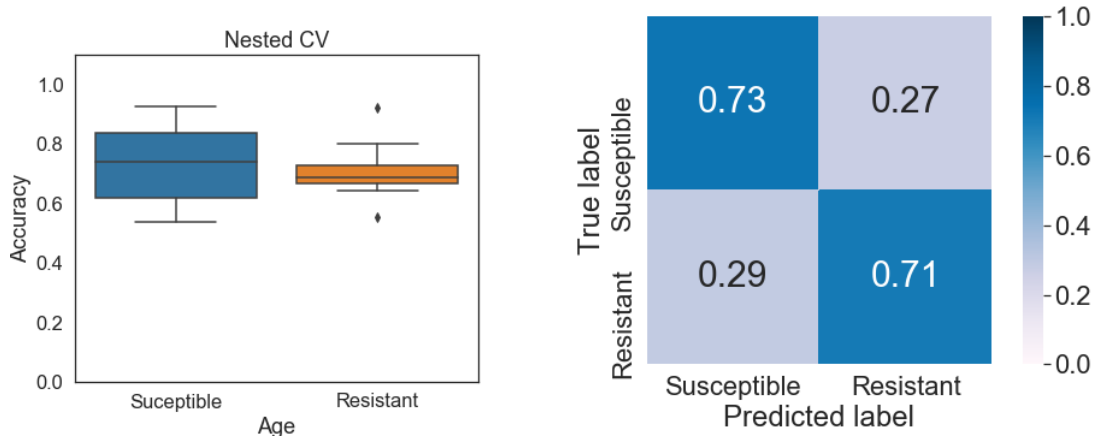


Figure 2.16: a) Boxplot of accuracies for each class: susceptible and resistant after ten-fold nested cross-validation. In boxplots, the horizontal line represent the median, boxes the interquartile range, whiskers the overall range, and black dots represent outliers. **b)** Confusion matrix showing the mean accuracy of 10 models after nested cross-validation

Table 2.2: Summary of the model accuracy for each of the classification problems using hold out set, nested cross-validation and bootstrapping

Classification Problem	Hold out	Nested-CV	Bootstrapping (n=100)
Species	80%	69.4%(8.8%)	66.0% (CI 95%: 59.2% , 71.5%)
Age	87%	77.1% (6.5%)	75.1% (CI 95%: 66.9% , 82.3%)
Insecticide resistance	69.77%	71.3% (8%)	69.6% (CI 95%: 62.7% , 77.2%)

Table 2.3: Summary of the processing time of traditional methods for age grading (parity status), morphological identification, NIRS, MIRS using ATR and µDRIFT

Method	Sample preparation/ Scanning time per mosquito	Accuracy	Ref
Parity dissection	8 min Detinova 45 min Polovodova	Gold standard for parity status	[293]
Morphological identification	Not possible in cryptic species	–	
NIRS	30 seconds	High accuracy for species prediction and parity status in *Anopheles*	[213, 218]
MIRS-ATR	less than a minute	High accuracy for species and parity status in *Anopheles*	[17, 199]
µDRIFT (legs)	30 sec - 45 sec	High accuracy for species prediction in *Aedes* High age prediction (3 and 10 days old) and low species prediction in *Anopheles*	[224]

2.4 Discussion

This study is an initial proof of concept on the potential of µDRIFT as a tool to identify biological traits in laboratory reared *Anopheles* mosquitoes. This technique allowed to collect spectra from the head, thorax, abdomen, and legs without destroying the sample. Spectra from one leg taken from 344 mosquitoes coupled and analysed by logistic regression allowed us to classify *An. coluzzii* and *An. gambiae* females with 62% accuracy, and 3 days and 10 days old with 89% accuracy.

This is the first attempt to identify two species of the *An. gambiae* complex, *An. gambiae* and *An. coluzzii* using µDRIFT. Analysis of the model coefficients showed that bands with more weight

in the model decision were mainly from the C-O stretch (940-1130 cm^{-1}), and C=O (1700 cm^{-1}) [292]. The C-O stretch region is related to chitin [292,294] while C=O band at 1700 cm^{-1} is assigned to protein and wax [17]. Evidence shows differences in survival in drier environments [295], cuticle hydrocarbons expression [296, 297] and cuticle thickness [298] between these two species. A previous study using ATR-FTIR for species classification of *An. arabiensis* and *An. gambiae* obtained 80% accuracy with laboratory colonies [17]. Moreover, μDRIFT has shown high classification accuracy between species of *Aedes* which are not cryptic [224]. However, there are no comparable studies using MIRS or NIRS to distinguish these same *Anopheles* species, therefore we cannot conclude that the low accuracy obtained here is due to the technique rather than the nature of the species. We hypothesised that some of the ecological differences between *An. coluzzii* and *An. gambiae* may not be associated with changes to the leg structure of sufficient magnitude to allow reliable differentiation, however, the classification power of our model may also be limited by the relatively low sample sizes.

My model correctly predicted the age of 3 days old and 10 days old with 87% accuracy and 77.1% ±6.5% depending on the evaluation method. My results are similar to those reported for *Ae. aegypti* where females were correctly predicted into age categories of 2 and 10 days with 81 to 88% accuracy respectively using Partial Least Squares Discriminant Analysis (PLS-DA) [222]. There are no studies in *An. gambiae* and MIRS predicting only two ages categories, therefore, comparing them to my results is not accurate. Gonzales et al. [17] were able to predict *An. gambiae* into age groups 3, 9 and 11 ranged from 20 to 50% and for *An. arabiensis* ranged from 40 to 80%. The accuracy of predicting mosquito age in broad categories of "young" (<7 days old) or "old" (7+days) using NIRS ranges from 65-97% [19, 196–198]. In NIRS analysis, there is often variability in prediction accuracy between groups from 3 to 9 or 12 days, depending on the algorithm used [198, 214]. The bands used by our model for age prediction were similar to those identified in two previous studies using MIRS [17, 222]. Like these previous studies, we identified important bands for age prediction to fall within the C-O stretching (940- 1130 cm^{-1}) and amide I region [292]. This C-O stretching region is assigned to chitin [292, 294], and the amide I region is assigned to proteins and N-H related to proteins. The CH_2 bend for chitin and wax were also important for age classification. It has been shown that gene expression changes with age in chitin bind cuticle proteins and chitin metabolism in *Anopheles* [206, 207]. Even though there was information in the literature about specific bands for age grading, at that time only one study showed feature importance for age grading in the infrared region [222]. Therefore, it was important to not bias the analysis to specific regions, since this was the first time mid-infrared was collected from the legs for age grading. Also, it was important to assess if the model would pick up signal from regions with meaningful biological information and if those bands match previous literature. These results put more evidence on the fact that age affects the cuticle structure of the mosquito. These structural changes are associated with the increase and decrease of cuticle proteins and also changes in the chitin metabolic process. Although these

changes as not been fully characterised, they can be detected with MIRS.

NIRS and MIRS have not been tested for identification of cuticular insecticide resistance in mosquitoes. Cuticular resistance is characterised by a thickening of the leg cuticle by the increased content of chitin, proteins and CHCs [179]. The top 10 variable contributions are located in the C-O stretching (940-1130 cm^{-1}), C-O-C ring (1130-1170 cm^{-1}) and Amide II (1500-1525 cm^{-1}) regions [294]. The C-O stretching and C-O-C ring, and Amide II associated with chitin and proteins. Cuticular resistant mosquitoes are characterised by a thicker procuticle at the tarsus, an increment in CHCs in the legs and a high content of chitin monomer D-glucosamine [179]. The differences may not only be in terms of chitin quantity but also in qualitative differences arising from down regulation of CPs, CPR8 and CPR120 [179]. These differences may have been identified by the algorithm. The question remains if this apparent difference can be pinpointed cuticular resistance, or is a confounding effect of strain (e.g., only 1 strain had cuticular resistance). I tried to disentangle the influence of the strains on the classification between “susceptible” and “resistant” by first assessing their differences using PCA, but we did not identify any strain specific clustering. This result can be interpreted as lack of evident differences in the spectra between strains. Moreover, when trying to classify between strains, just one strain, Tiassale, was identified with a high accuracy. The model was not able to predict Ngousso, with accuracy below 50% and slightly above 60% for Kisumu. I expected that if the strains were different, the model should be able to differentiate between the three with at least 60%. I thought that the approach "one vs the rest" may influence the classification but I tested also "one vs one" with similar results. Future work needs to focus to disentangle the strain/cuticular resistance confounder. One way can be the analyses of multiple populations of the same species and strains with different degrees of cuticular resistance and asses how model classify them. Another option is to measure the cuticle thickness (by electron microscope) and chitin content from the legs and asses if they are correlated at an individual leg mid infrared spectrum. Moreover, sample size needs to be increase.

Our models were evaluated using hold-out data set to avoid any over or under optimistic results; but we found upwardly accuracies in all our classification problems (Table 2.2). This was especially high when predicting age, which went from 87% using validation down to 77.1% with nested cross validation. This phenomenon was less dramatic in models classifying species and cuticular resistance. Optimistically biased performance estimates can be linked to the small sample size used, therefore, nested cross-validation and bootstrapping were also implemented. Nested cross-validation is known to give unbiased performance estimates regardless of sample size [288, 299]. Nested cross-validation also gave us the ability to report accuracy values with a standard deviation, which provides a more accurate representation of the performance of the models. Future work should be performed with a larger sample size to confirm the accuracies reported here.

Pre-processing was also tested to see how affect model accuracy. It is widely use in NIRS studies to

correct light scattering effects and baseline shifts. The aim is to maximise the difference between classes while minimising the differences between samples of the same group. Pre-processing in MIRS is not as important as NIRS mainly because of the use of ATR, which minimises baseline shifts and scattering problems in MIRS in contrast to reflection in NIRS. Pre-processing in studies using ATR varies widely, from mean centering, to use of first and second derivatives. Savitzky-Golay second derivative improved model performance in species identification. This agrees with studies in *Aedes* and MIRS [222,224]. However, for age prediction, it did not generate a difference in model performance, except when using second derivative and 21 windows smoothing together with LR, where model accuracy dropped to 50%. This might be due to the nature of logistic regression, since the rest of the models did not show this behaviour. Moreover, the extreme window smoothing might eliminate any differences between spectra.

µDRIFT scanning time was different for tick tissues (head, thorax, and abdomen) and for legs. Long scanning times, up to 5 minutes, were required for tick tissues. This scanning time is higher compared to scanning speeds reported in ATR spectroscopy and NIRS. However, the legs only required 16 scans with a resolution of 4 cm^{-1} which is less than 10 seconds. As mentioned in the methods section, a single mosquito needs approximately 35 seconds from removing the legs until collecting its spectrum. This scanning time is more efficient than parity dissection (8 minutes per mosquito Detinova method to 45 minutes Polovodova method [293]) or morphological identification (which cannot be done with cryptic species). Moreover, scanning time can be reduced even further by collecting only the region from 1800 to 600 cm^{-1} to match NIRS scanning time, which is 3 to 5 seconds. WHO insecticide susceptibility test requires 1 hour of exposition plus 24 hours to measure mortality. At 20-25 samples per tube with six replicates, it can process 150 female mosquitoes per test, each mosquito requires 10 minutes. This time can be reduced, but it will require more test tubes and manpower, while spectroscopy just need one operator and the spectrometer. In terms of processing time, µDRIFT is at the same level of NIRS and MIRS using ATR (Table 2.3).

Diffuse reflectance spectroscopy is very sensitive to changes in the path length, and one difficult encountered here was positioning mosquito legs as flat possible over the gold mirror to avoid distorted spectra from the samples. The leg joints can interfere and create a gap between the tissue and the gold mirror that generates changes in the spectra; an issue previously raised in Sroute et al. [224]. We tried to minimise this problem by separating the femur from the rest of the leg, and eliminating the joints of the leg and eliminating the joints to avoid positioning the sample incorrectly. This process increased the overall processing time at the beginning, but it can be done easily with training. Future work should focus on analysis of other species (i.e., *An. arabiensis, An. funestus*) to assess whether the low accuracy in species discrimination reported here is due to the close genetic relatedness of the species used here or is a general feature of the approach. Also, further study on in mosquito groups expanded to encompass differences in terms of feeding and physiological status is needed to increase the robustness of the technique.

There are several limitations in this chapter. First, resistance/susceptibility was not validated by traditional techniques. Therefore, it is not sure if the strains were 100% resistant or susceptible. This is also true for species. However, routine species identification is performed every month to assess the purity of the colonies reared in the laboratory. Future work should include validation by PCR for species identification and WHO bioassays for cuticular resistance. Second, the decision of measuring only the legs for species identification/age grading was due to the longer scanning times required for thick tissues (head, thorax, abdomen) compared to the legs. I acknowledge that literature co-relate chemical information with age in the head and thorax. However, legs have been also used for age relating studies [204, 205, 293] and also emerging literature was indicating the legs as a potential tissue for MIRS in species identification [224]. On top of that, cuticular resistance is localised in the legs. All these reasons made the case to focus on the legs as an interesting body part to assess with μDRIFT in this chapter. However, the rest of the tissues should be considered and explore in detail to take advantage from the chemical information in the thorax and the head.

Diffuse reflectance spectroscopy offers a rapid non-destructive method for measurement of mid infrared spectra from mosquitoes for species, age, and insecticide resistant classification. The advantage of using the legs is the ability to scan numerous samples at the same time by placing them next to each other, facilitating the scanning process. One of the advantages compared to ATR spectroscopy is the non-destructive nature of the process, which leaves the remaining mosquito specimen available for further analysis (e.g., PCR or ELISA analysis for infection, etc.). It allows exploring tissues such as wings and legs which might serve to identify other biological traits, not only in mosquitoes but other diseases vectors. The accuracy is comparable in terms of age grading. Nevertheless, the added cost and the size of the microscope can be a limiting factor in terms of cost ($ 25000 for the added microscope). However, new IR microscopes with Quantum Cascade Lasers are becoming cheaper and smaller than current FTIR systems; therefore, in the future these systems will become easier to implement and test them. Compared to NIRS, both techniques are non-destructive but NIRS have been tested longer with high accuracy for age grading, species prediction and protocols have been standardised. The proposed technique here is in its infancy and at the moment is not recommended as an alternative to NIRS. Having said that, diffuse reflection opens the door for non-invasive and targeted measurements that can expand vector surveillance tools. On top of that, new technologies (i.e., quantum cascade lasers) have shown potential to overcome the limitations of diffuse reflection measurements. Currently, they have been used on human dermis for non-invasive glucose monitoring [54, 110]. This technique is a promising approach despite current limitations, and it is worth exploring its strengths with further experiments

2.5 Conclusion

This study adds to the growing evidence on the value of MIRS coupled with machine learning algorithms as a tool for vector surveillance of mosquitoes of medical importance, specifically malaria vectors. Moreover, it shows the potential of diffuse reflectance spectroscopy for age grading and species identification. In summary: i) Spectra from the legs can be acquired significantly faster than those from other mosquito parts (head, thorax, abdomen). ii) Spectra from one mosquito leg is sufficient for use in classification problems. iii) Logistic Regression performed relatively well, predicting mosquito age. iv) Even with limited sample size, the model performance for identification of species and a cuticular resistant strain was over 60%. v) Nested cross-validation gave unbiased accuracy across all classification problems. Larger data sets for training testing and validation will likely increase model performance. Moreover, inclusion of different species (*An. funestus* and *An. arabiensis*) and group ages (from 1 to 15 days) is needed to further test the robustness and applicability of this technique. It may also be useful to include more accurate feature selection and reduce bands in order to increase the performance of the models. The first step to illustrate that a cuticular resistant strain can be identified with MIRS has been taken, but a lot of uncertainty remains given the potential confounding of resistance phenotypes and correlated strain attributes in this study. The use of the same strain with different degrees of cuticular resistant, multiple strains of the same species with known differences in resistance, measure of leg thickness and chitin content can help disentangle resistance phenotype and strains. This study shows the potential of μDRIFT using the mid-infrared region for mosquito surveillance. Future work will determine if the technique can predict insecticide resistance as well other species and more age groups

Chapter 3

Species Classification and age grading of *Anopheles* mosquitoes using multivariate and machine learning models

3.1 Introduction

Mid infrared spectroscopy (MIRS) has been recently explored as a complementary technique to Near Infrared Spectroscopy (NIRS) in mosquito species prediction and age grading [17, 222]. MIRS bands are due to discrete fundamental vibrations of biomolecules providing more information about the chemical composition of a sample, compared to the broad bands of NIRS. This might be useful in detecting small changes in the cuticle due to age, or between species. The most used analytical technique is Attenuate Total Reflection - Fourier Transform infrared (ATR-FTIR) combined with supervised nonlinear machine learning algorithms. Similar to all spectroscopy techniques, ATR-FTIR is not immune to artefacts in the spectral signal. These artefacts can be grouped into scattering effects and baseline offset/slope [290, 300]. Scattering effects are caused when the size of the particles in the sample has one dimension that is roughly the same as the wavelengths used [300, 301]. Baseline offset is another typical characteristic of ATR-FTIR, where the baseline is lifted by a constant value. Slope baseline is when baseline is lifted by no constant values, which increase slowly. Other effects can be attributed to low signal intensity, etc. (Fig. 3.1)

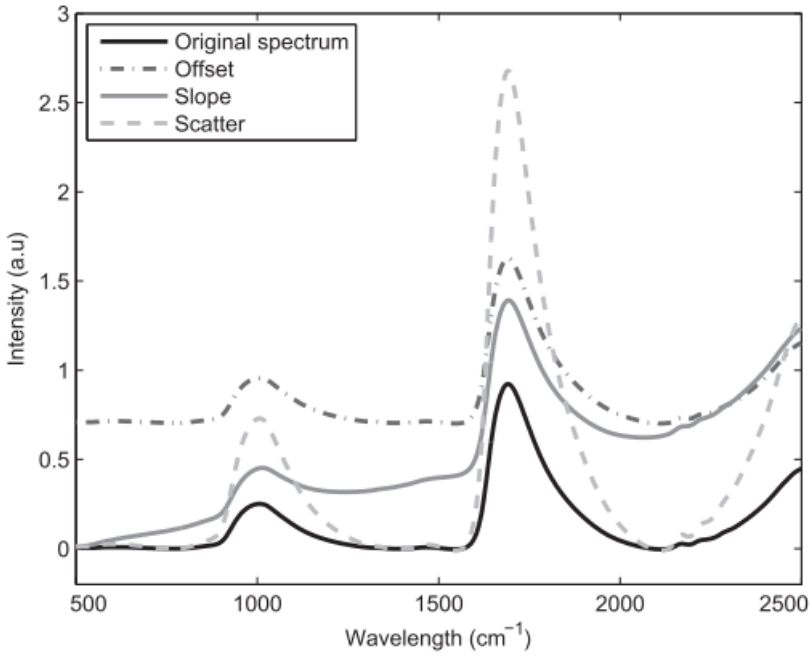


Figure 3.1: Representation of artefacts in spectral data. Black line is the original spectrum, follow by version affected by artefacts. Modified from Jansen, J. et al., 2013

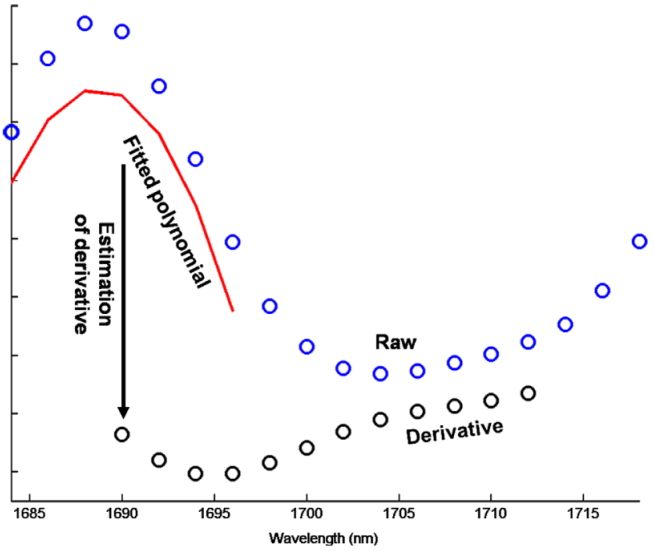
After the collection of quality data, pre-processing is one of the most important steps before using Principal Components Analysis, Partial Least Squares, or any other machine learning algorithms to analyse spectra [290]. Pre- processing allows correction of scattering effects, baseline slopes and offsets, and any other artefacts that might interfere with the performance of classification and regression models. Here, we are going to describe some of the most common pre-processing algorithms.

3.1.1 Types of pre-processing and scatter correction

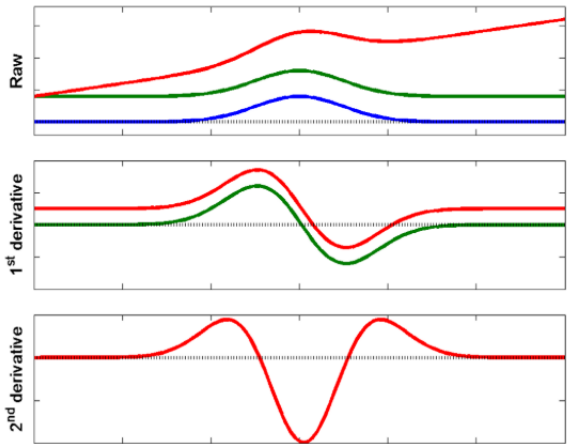
Savitzky-Golay

The most common smoothing algorithm is the Savitzky-Golay (SG) algorithm. SG fits a polynomial (of degree dg) in a small symmetric window of at least $dg + 1$ ($dg = 2m - 1$) points that is centred in the point to be replaced by the smoothing value (one polynomial per window) [290]. Once the data point has been replaced, the window shifts one data point, and the process is repeated (Fig. 3.2a). The main advantage of the SG is that it preserves the relative maxima, minima, and width of the spectral data while reducing any random noise. However, wider filter windows can distort the signal. Therefore, the filter window, as well as the polynomial degree, need to be optimised [302].

To enhance spectral resolution and eliminate baseline, SG derivatives can be used for differentia-



(a)



(b)

Figure 3.2: a) Estimation of first derivative using Savitzky-Golay with a 7 window. **b)** Effect of first and second derivative on raw spectra (blue), spectrum with additive effects (green) and additive plus multiplicative effects (red). Modified from Rinnan, A. et al., 2009

tion. The first derivative removes additive constant background effects, and the second derivative removes the baseline linear slope variations and additive effects (Fig. 3.2b). The number of points used to calculate the polynomial and the degree of the fitted polynomial can be chosen according to how much noise the raw signal has [303, 304]. The first and second derivatives may help identify the less apparent features, but a larger filter window size might remove relevant information [302].

Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) removes scatter effect artefacts or imperfections from the spectral data. This process also accounts for scaling baseline effects [303]. It requires a reference

spectrum that should not have any scattering effects. However, reference spectra are usually not available for every type of sample, thus, the mean spectra can be used as a reference spectrum. MSC is based under two assumptions: First, a spectrum is composed of two spectra: one due to light diffusion (d) and the other due to chemical absorbance (c), where the spectrum d needs to be corrected:

$$x_i = d_i + c_i \tag{3.1}$$

Second, all samples have the same coefficients of the diffusion spectrum at all wavelengths. This allows the estimation of the correction coefficients, b_o and $b_{ref,1}$ using ordinary least-squares regression of the individual spectrum against the mean spectrum [305] using the following model:

$$X_{org} = b_o + b_{ref,1}.X_{ref} + e \tag{3.2}$$

and the final corrected spectrum is obtained by subtracting:

$$X_{corr} - \frac{X_{org} - b_o}{b_{ref,1}} = X_{ref} + \frac{e}{b_{ref,1}} \tag{3.3}$$

where X_{org} is the uncorrected spectrum, X_{ref} is a reference spectrum or mean spectrum, e is the un-modeled part of X_{org}, X_{corr} is the final-corrected spectrum, and b_o and $b_{ref,1}$ are correction coefficients.

Each spectrum from the data is shifted and rotated to fit the reference spectrum. After correction, all spectra will have the same scatter level as the reference. To calculate b_o and $b_{ref,1}$, a sub-region from the reference with no chemical information is selected, which works well if the reference spectrum does not have any variation. The use of the mean spectrum as reference might remove information and decrease the performance of model predictions. The success of MSC relies on finding the best reference spectrum, however, when using the mean spectrum of the calibration set, it might not represent new data well.

Standard Normal Variate

Standard Normal Variate (SNV) is also one of the most common pre-processing algorithms used on NIR data. It was introduced by Barnes *et al.* [306] and reduces the multiplicative effects of scattering and particle size, and reduce the differences in intensities of the signals [302]. Contrary to MSC, it does not require a reference spectrum. To correct spectral data suing SNV, each spectrum is centred and then scaled by dividing by its standard deviation using the following equation:

$$X_{corr} = \frac{X_{org} - a_0}{a_1} \tag{3.4}$$

where: a_o is the mean value of the spectrum to be corrected and a_1 is the standard deviation

This transformation is carried out on individual samples and not to the whole data set as MSC. SNV can improve Partial Least Squares model prediction when spectral data are affected with scattering effects by reducing within-class variance; however, the corrected spectra will have positive and negative values centred on 0, making interpretation more difficult. Finally, artefacts could be introduced when applying SNV. his is caused by the assumption of SNV is that multiplicative effects are uniform over the whole spectrum, which is not always true [302].

Robust Normal Variate

Robust Normal Variate (RNV) was developed to reduce the artefacts introduced by SNV by using the median or the mean of the inner quartile range and the standard deviation. First proposed by Guo *et al.* [307], he defined the problem with SNV as the "closure" which is defined as "a statistical term indicating that the sum of the data is necessarily equal to a certain amount, so that if one of the data changes in one direction, the other data must change into the other direction to compensate for the change". Therefore, when SNV is applied, the data are closed due to the sum of the deviation of each spectral values from the mean absorbance is zero. This phenomenon introduces artificial negative correlation. RNV eliminates closure by subtracting the median of the spectrum from each spectral variable and divide that value by the standard deviation of the spectrum [303]:

$$z = \frac{[X - percentile(X)]}{std[x < percentile(x)]} \tag{3.5}$$

where $std[x \leq percentile(x)]$ means the standard deviation of values that are less than the percentile. RNV is less sensitive to outliers compared to SNV; however, optimisation of the percentile level is needed.

3.1.2 Extracting information from spectral data

After the signal is pre-processed, the next step is to extract meaningful information from the data to build models that can predict or estimate information when presented with new samples. Although there is an extensive list of algorithms for that purpose, in this chapter, I am going to describe the most widely used in chemometrics, partial least squares, and partial least squares discriminant analysis.

Partial Least Squares regression

Partial Least Squares (PLS) regression is a general method for path analyses using latent variables [308]. It assumes that both the independent matrix X, (the absorbance values for each wavelength) and the dependent matrix Y (the species names or age days) can be projected onto

a low-dimensional factor space and that a linear relation exists between the scores of the two blocks [309]. The dependent and independent matrices are decomposed into scores and loadings by:

$$X = TP^T + E_X Y = UQ^T + E_Y \tag{3.6}$$

where T, U, P, and Q are the X and Y scores and loadings matrices respectively. A linear relationship between X and Y scores is given by:

$$U = TC \tag{3.7}$$

where C is the diagonal matrix of coefficients. Then it is possible to compute the regression coefficients B which will be used to predict the values of the matrix Y for unknown samples from which X is known.

$$Y_{new} = X_{new}B \tag{3.8}$$

Partial Least Squares Discriminant Analysis

Partial Least Squares Discriminant Analysis (PLS-DA) is a multivariate dimensionality reduction method. It is a modification of PLS where a binary coded dummy variable, Y, is created based on the class information of the training data set. Then the regression model is calculated based on the PLS algorithm [309].

3.1.3 Traditional chemometrics applied to species and age classification in *Anopheles* mosquitoes using MIRS

The selection of pre-processing methods in ATR-FTIR data has not been given considerable attention [300]; with mosquito studies using ATR-FTIR being no exception. Out of the two studies using ATR-FTIR method [17, 222], only one uses Savitzky-Golay. Moreover, studies of species classification and age grading in *Anopheles* have not used the PLS algorithm so far. The arguments against using PLS are: first, its propensity for over-fitting; meaning the model adapts "too well" to the training data and has limited value for predicting unseen data [17]. This phenomenon has been seen in NIRS [214], where PLS was not able to predict ages of field samples reared to known ages. However, the other models used in that study (Oblique Random Forest and support vector machine) suffered from the same issue. Therefore, overfitting might not be the problem, but the fundamental differences between training data and the unseen data. This issue might not be resolved by more complex models. There is evidence of the use of encoders and ANN to increase model generalisation, but it has only been used in NIRS [218]. It is worth investigating if MIRS also suffers from the same issue, and it is due to the data or the model itself. The second argument is that PLS assumes linearity, which for species prediction does not represent an issue.

This is seen in the high accuracy for species prediction in NIRS and MIRS. However, for age grading, it represents a problem, since changes in the cuticle caused by age might not be linear. This has been seen in NIRS studies where PLS underestimates older mosquitoes when trying to predict chronological age [214]. Consequently, there has been a preference for using non-linear supervised machine learning algorithms in the hope of overcoming the limitations of PLS. Recent studies have compared PLS and ANN with an increment of accuracy using ANN for age grading [231]. Currently, there has been no proper assessment of the prediction power of PLS and PLS-DA with mid-IR data, nor a comparison of PLS with other machine learning models. Therefore, the objectives of this chapter are to:

1. Establish a baseline using PLS-DA and PLS for Anopheles mosquito species classification and age prediction using ATR-FTIR data.
2. Assess pre-processing algorithms for ATR infrared data and how they affect model accuracy in species classification with linear and nonlinear machine learning models.
3. Identify the region most informative of the infrared spectra for *Anopheles* species classification

3.2 Methods

3.2.1 Data sets

Three datasets were used for this chapter. Data set one was obtained by researchers from the Chemistry department of the University of Glasgow. The specimens from University of Glasgow are female laboratory reared mosquitoes of two species: *Anopheles gambiae s.s* (Kisumu strain) and *An. coluzzii* (Ifakara strain). Rearing, blood feeding, and processing of mosquito specimens are described in [17]. Briefly, the mosquitoes were reared under standard insectary conditions of 27 ± 1 C, 70% humidity and a 12-hr light: 12-hr dark cycle. *Anopheles gambiae s.s* (Kisumu strain) mosquitoes were provided by Prof. Hilary Ranson (Liverpool School of Tropical Medicine). Larvae were fed ad libitum on fish pellets (Tetra Pond Pellets, Tetra GmbH, Herrenteich 78, D49324). Pupae were collected from larval rearing trays and moved into a cage for adult emergence. Mosquitoes were considered to be in the age category of "Day 0" on their day of emergence from pupa to adult. Upon emergence, adults were fed *ad libitum* on a 5% glucose solution. Mosquitoes were held from 1 up to 15 days post emergence (with an interval of one day) for analysis. Upon collection, mosquitoes were transferred into a cup and killed with a cotton soaked with chloroform placed on top of the cup for 30 minutes. Dead mosquitoes were then transferred into a tube over a layer of cotton and silica gel desiccant. The vial was then immediately

stored at 4°C for at least 3 days before scanning. Spectral acquisition from individual mosquito thoraxes was taken by Attenuate Total Reflection FTIR using a dry-air purged Bruker Vertex 70 spectrometer equipped with a globar lamp, a DLaTGS detector, a KBr beamsplitter, and a diamond ATR accessory (Bruker Platinum ATR Unit A225). Sixteen scans were taken at room temperature between 400 and 4,000 cm^{-1} with 1 cm^{-1} resolution. Datasets two and three were obtained by researchers at the Institut de Recherché en Sciences de la Santé (IRSS), Burkina Faso. The laboratory colonies reared at IRSS (IRSSlab) are *An. gambiae*/Soumousso and *An. coluzzii*/ Vallée du Kou and were reared at laboratory conditions as described above. Mosquitoes were sampled at ages ranging from 1 to 17 days old, with a hundred and twenty mosquitoes per day. Each age group is composed of three physiological status. The IRSSfield dataset contained the same laboratory colonies but reared under semi-field conditions. These laboratory colonies were reared during the rainy season. Temperature and humidity were constantly monitored every day in the semi-field and recorded. Mosquitoes blood fed from live cattle and were able to oviposit in water containers inside the cages. The collected age groups were 1, 4, 7, 10, and 15 with 50 mosquitoes per age group [310] (Table 3.1)

Table 3.1: Datasets information of origin, institute, number of samples, rearing conditions and species

Dataset	Origin	Institue	# Samples	Conditions	Species	Age groups
Glasgow lab	Glasgow	University of Glasgow	4446	Laboratory	*An. coluzzii* *An. gambiae*	1d to 17d
IRSS lab	Burkina Fasso	Institut de Recherche en Sciences de la Santé (IRSS)	4926	Laboratory	*An. coluzzii* *An. gambiae*	1d to 17d
IRSS field	Burkina Fasso	Institut de Recherche en Sciences de la Santé (IRSS)	522	Semifield	*An. coluzzii* *An. gambiae*	1d,4d,7d,10d,15d

3.2.2 Data pre-processing

Data was pre-processed using Python 3.6 with in-house developed scripts. Assembly of datasets from individual files and cleaning (low intensity, high water content, atmospheric intrusion) was done using Loco Mosquito 5.0, a program written in Python 3.6 [17]. The spectral region used was from 1800 to 600 cm^{-1}, except when stated differently. Before data analysis the following pre-processing algorithms were applied on the spectra: Standard Normal Variate (SNV), Multiplicative Signal Correction (MSC), Robust Normal Variate (RNV) [291] and Savitzky-Golay.

3.2.3 Statistical Analysis

Partial Least Squares regression (PLS) and five machine learning models: Logistic Regression (LR), Support Vector Machines (SVM), Decision Tress (CART), Support Vector Machine with Stochastic Gradient Descent training (SGD) Gaussian Naives Bayes (NB) were tested using scikit-learn package [289]. Absorbance at different wavenumber values were used as features for model training.

Partial Least Squares analysis

To establish a baseline for species classification and age grading, PLS-DA and PLS regression were used to predict species and estimate mosquito age, respectively. Glasgowlab dataset and IRSSlab dataset were used to optimise the number of latent variables via cross-validation for species and age. A linear search was used to optimise the number of latent variables (from 1 to 400). First, the number of components with the lowest root mean squared error (RMSE) were chosen with only Glasgowlab data set (Appendix Fig B.1). Second, to avoid overfitting, meaning choosing a model with numerous components which adjust too closely to the training data set, a new linear search from 1 to the number of components already chosen was performed using IRSSlab dataset as an independent test set. Area Under The Curve (AUC) of the Receiver Operating Characteristics (ROC) was used as a metric (Appendix Fig. B.2). The number of latent variables that balanced high values of AUC-ROC when using both Glasgowlab and IRSSlab data set was chosen. After optimisation, the Glasgowlab data was split into two sets: training and test sets (50% each). The training data set was split further into training and validation sets (70%, 30%). The model was trained, tested and then validated. This process was performed 50 times for each scatter correction and also using IRSSlab and IRSSfield as validation datasets. The model with the highest AUC-ROC was chosen for the final confusion matrix and density plots. For age classification, Glasgowlab dataset used for optimisation of the number of latent variables using cross-validation. The R^2 value was used to assess the best model. Validation was also evaluated with IRSSlab and IRSSfield. Model coefficient values were used for wavelength importance for species and age. Means for each day were calculated and analysed using an analysis of variance (ANOVA) test with Tukey's correction for multiple comparisons.

3.2.4 Evaluation of scatter correction algorithms on machine learning algorithms

The process of evaluation was the same as described for PLS. Accuracy value and confusion matrices were used to assess the performance of the model.

3.2.5 Prediction accuracy of different spectra regions

Further analysis to identify which regions contained more information for classification purposes was done using three main spectral "windows": 950-1250 cm^{-1}, 1250-1500 cm^{-1} and 1500 - 1800 cm^{-1}. Each window was based on the range of quantum cascade lasers. The region of 950 - 1250 cm^{-1} is based on the current range of our QCL setup (Chapter 4). For each window, cross-validation was used to calculate baseline accuracy. Then, each region was evaluated for species prediction IRSSlab and IRSSfield datasets. After that, the window with the highest accuracy was split into three regions. The process of training and evaluation was repeated. Lastly, to explore the spectral differences between the two species, prediction coefficients from the best models were extracted and plotted against wavenumber values. The dominant wavenumber values were used to infer the possible biochemical difference of each species.

3.2.6 Analysis of Spectra

Principal component analysis (PCA) of all combined data sets was performed to visualise any clustering of the samples. Loading plots of principal components with the most explained variance were analysed to identify wavelength differences between data sets.

3.3 Results

A total of 9894 mosquitoes were used in the analysis (Table 3.2). The spectral region used was 1800 to 600 cm^{-1} except where otherwise indicated.

Table 3.2: Number of samples used in the analysis

Data set	# *An. gambiae*	# *An. coluzzii*	Total
University of Glasgow	2019	2337	4446
IRSS lab	2000	2992	4926
IRSS field	264	258	522
			9894

3.3.1 Pre-processing and Scatter correction

An overview of how the different scatter corrections and pre-processing algorithms affect the shape of the spectra is shown in Fig. 3.3. Raw data show some scatter effects across the spectra, especially additive effects. After scatter corrections, a visible decrease of variation between sample can be observed, especially in the region 1500 to 1200 cm $^{-1}$. As expected, RNV produces fewer artefacts compared to SNV. The reason is the use of a percentile instead of the mean. SG with 9 point window size and second derivative eliminates any additive effect present in the data, as well as enhances the bands located in the 1700 and 1200 cm^{-1} region.

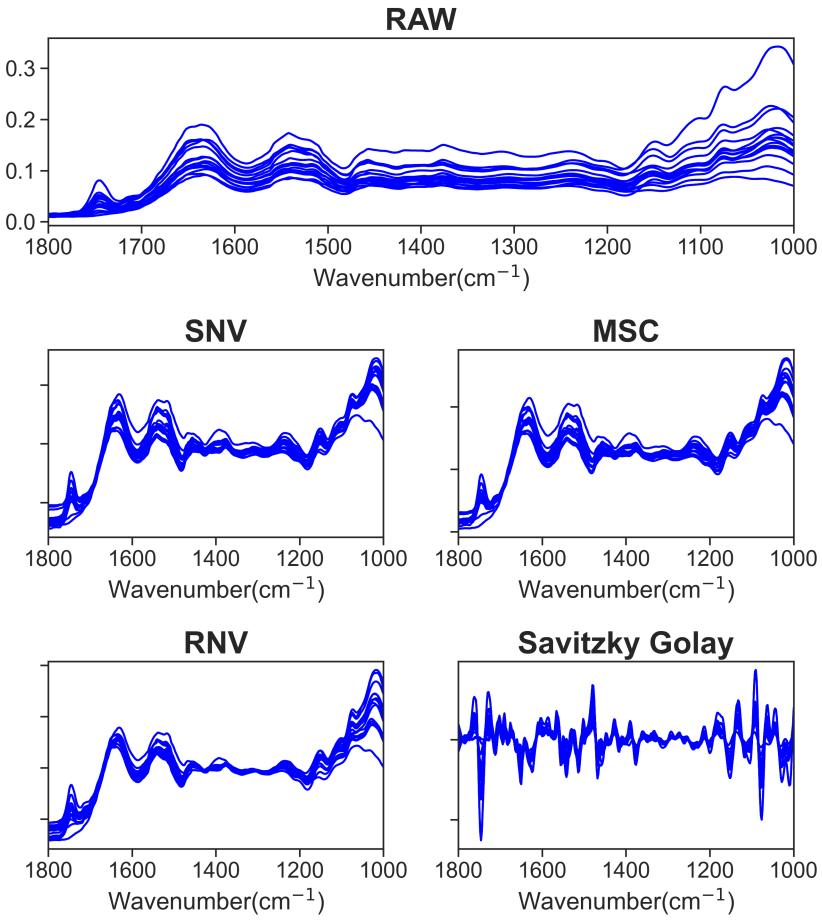


Figure 3.3: Effect of pre-processing on ATR-FTIR spectral data Mid infrared spectra from all laboratory reared female mosquitoes collected using ATR-FTIR in Glasgow without pre-processing (RAW) and after application of scatter correction algorithms, Standard Normal Variate (SNV), Multiplicative Signal Correction (MSC), Robust Normal Variate (RNV) and Savitzky-Golay on the same specimens. Each line represents the spectrum of one mosquito.

3.3.2 Classification of species using Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA models using MIRS can classify between laboratory reared females *An. gambiae* and *An. coluzzii* with high accuracy (Table 3.3). A model with 5 latent variables showed a high accuracy for species classification when laboratory reared mosquitoes from the same laboratory as test set was used (Fig. 3.4a-c, AUC = 0.93, Accuracy=0.87). Probability distribution plots showed a separation between classes and a high true positives and true negative values (Fig. 3.4c). The PLS model performance on samples from the same origin is excellent. However, its classification power decreased when independent validation sets were used. Model accuracy decreased when predicting mosquitoes from other laboratories (Fig. 3.4c, AUC=0.63, Accuracy=0.62). Probability density plots showed how the classes overlap with each other and the confusion matrix indicates how *An. coluzzii* can be classified (Fig. 3.4e, 3.4f). For IRSSfield, the model was unable to classify either species with AUC values below a random model (Fig 3.4g, AUC=0.35, Accuracy=0.47). This can be seen in more detail in the density plots and confusion matrix, where the model classifies *An. gambiae* more often as *An. coluzzii* and vice versa (Fig 3.4h and 3.4i).

Scattering correction and pre-processing increased model AUC-ROC values tested on samples from the same origin, but when tested with independent validation sets, there was a decrease in AUC-ROC. RNV was the exception with the highest AUC-ROC value of 0.41 when tested with the IRSSfield data set. The Savitzky-Golay second derivative had a detrimental effect on the model. It reduced AUC values further from 0.61 to 0.56 in the IRSSlab data, set and from 0.37 to 0.28 for IRSSfield (Table 3.3).

Table 3.3: Pre-processing, latent variables (LV) and Area under the curve of the Receiver operating characteristic curve (AUC) for each data set: Glasgowlab, IRSSlab and IRSSfield

Pre-processing	LV	AUC Glasgow	AUC IRSSlab	AUC IRSSfield
RAW	5	0.93	0.63	0.35
SNV	5	0.95	0.62	0.40
MSC	7	0.95	0.61	0.37
RNV	6	0.95	0.60	0.41
Savitzky-Golay	6	1.00	0.56	0.28

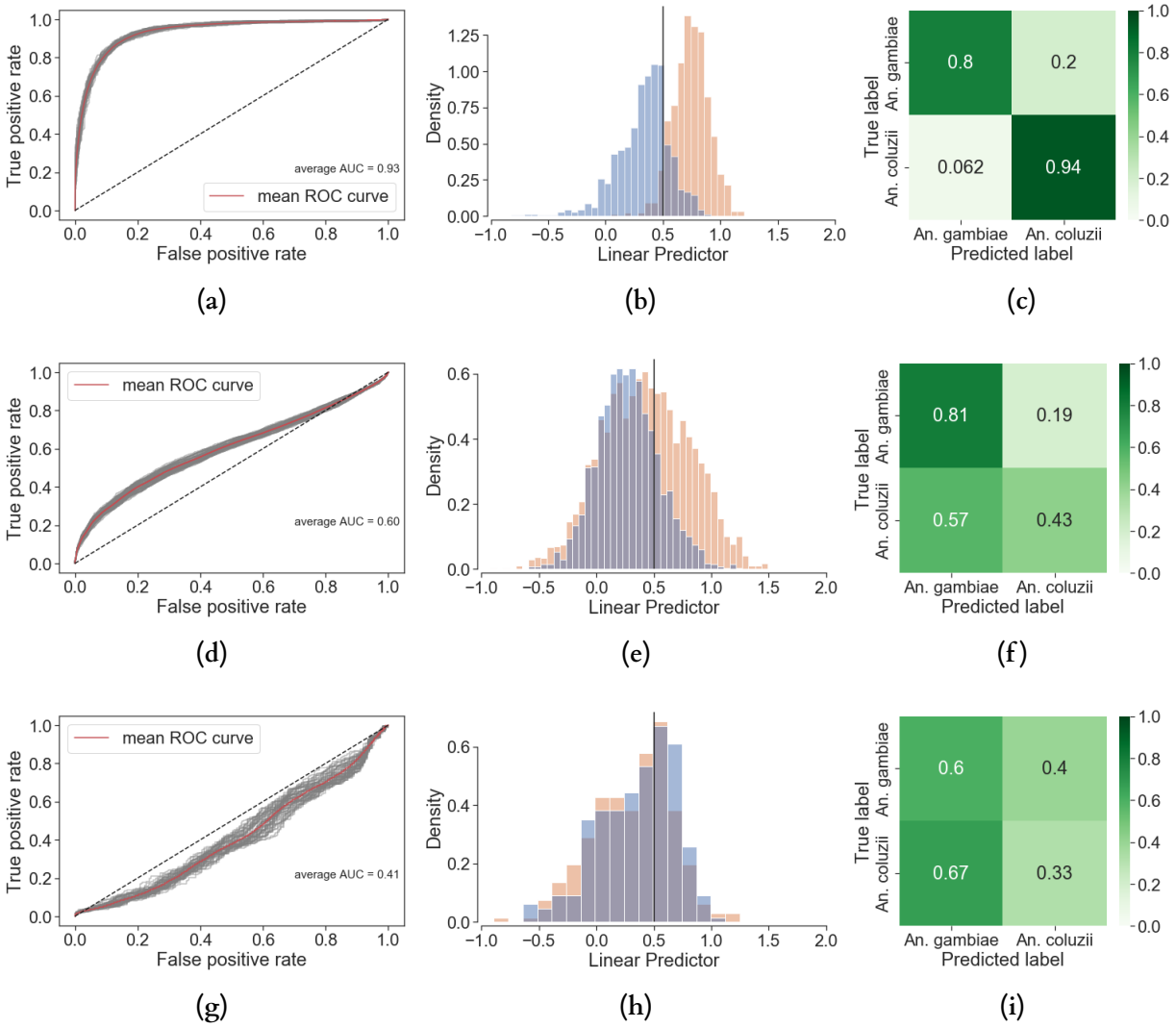


Figure 3.4: Species prediction using PLS-DA Prediction of *An. gambiae* (blue) and *An. coluzzii* (orange) using PLS-DA and tested with three different validation data sets. a) Receiver operating characteristic curve (ROC) of 50 PLS-DA models and mean corresponding Area Under the ROC curve using laboratory reared mosquitoes from Glasgowlab as validation set. b) Ability of the model to predict *An. gambiae* and *An. coluzzii*. Histogram of the estimated linear predictor for the validation set. Colour coded as follows: *An. gambiae*(blue) and *An. coluzzii* (orange). The vertical black line indicates the threshold for classifying mosquitoes into the two different species. Areas where distributions overlap indicates misclassified observations. c) Normalised confusion matrix for the best model on validation set. Each row represents instances of true class, while each column represents instances of predicted class. The same results are shown using mosquitoes from IRSSlab (d, e, f) and IRSSfield (g, h, i).

3.3.3 Age prediction using Partial Least Squares

MIRS with PLS regression models can predict moderately the chronological days of laboratory reared mosquitoes from 1 to 17 days old (Fig. 3.5, R^2 = 0.68, RMSE = 2.24). Differences between

very young (1-4 days old) and very old (12-17 days old) can be observed. The model overestimated the age of females in 1 to 4 days range, and underestimated the age of older mosquitoes from 14 to 17 days old (Appendix B, Table B.1)

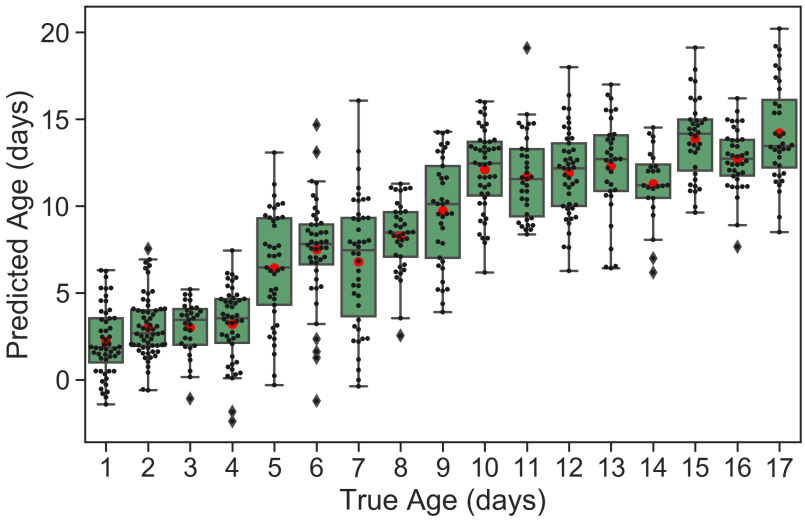


Figure 3.5: Prediction of chronological age Comparison of predicted vs true age tested on hold out set using PLS across age groups from 1 to 17 days old. Data is from Glasgowlab dataset. Each point represents a population sample. Boxplots show the upper and lower ends of the centre box to indicate the 75th and 25th percentiles. The red dot represents the mean, the line inside the box indicates the median and the whiskers show the maximum and minimum no further than 1.5 Interquartile Range (IQR). Outliers and indicated as a diamond.

After the first assessment of age prediction with all age groups, 6 ages were chosen to see if by increasing the number of days between each group, the model improves its ability of age prediction. The ages were 3, 5, 7, 9, 12, 15 days old. Predicted age using the raw and pre-processed data are shown in Fig 3.6. Savitzky-Golay tends to reduce the under and overestimation of ages (Table 3.4). Tukey Post hoc analysis indicated that *An. gambiae* could be differentiated into 5 age groups (3, 5-7, 9, 12, 15) when using raw data. After applying Savitzky-Golay pre-processing, 5 and 7 days could be differentiated into two different age groups (Fig. 3.7 a, b). The reason is that Savitzky-Golay first smooths any difference between samples of the same age, and then the second derivative reduces further differences by removing baseline differences, making the samples more homogeneous. These results on better age group separation (Fig. 3.7 b).

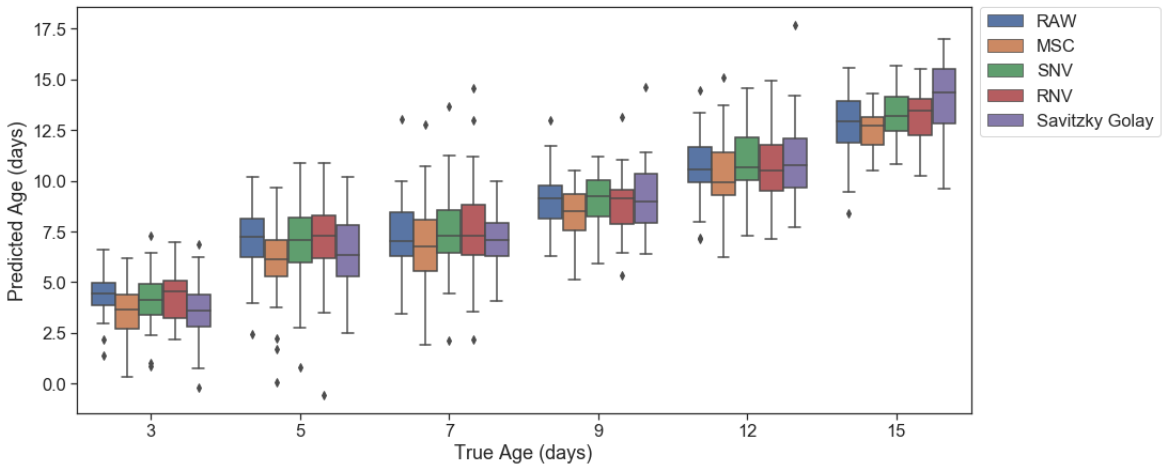


Figure 3.6: Effect of pre-processing on age prediction Comparison of predicted vs true age with 4 pre-processing methods: MSC, SNV, RNV and Savitzky Golay (window = 9, second derivative) using PLS. Model was trained with Glasgowlab data set and tested with a hold out set. Boxplot shows the upper and lower ends of the centre box to indicate the 75th and 25th percentiles. The line inside the box indicates the median, and the whiskers show the maximum and minimum no further than 1.5 Interquartile Range (IQR). Outliers are represented with a black diamond.

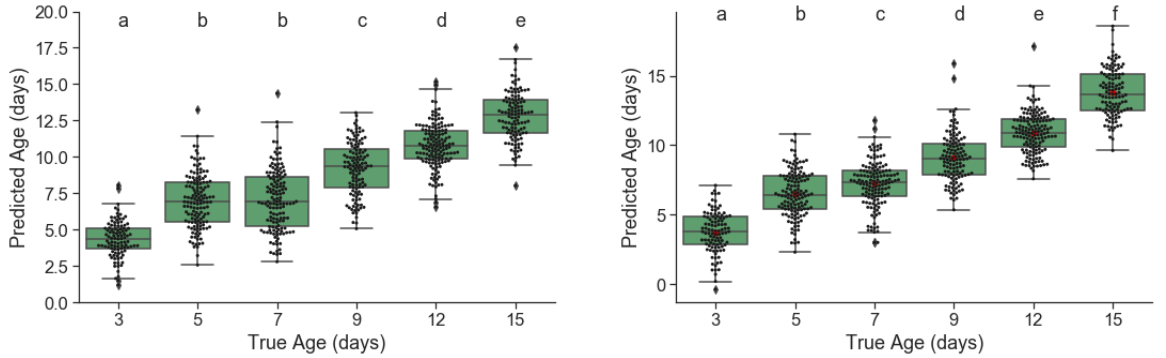


Figure 3.7: Effect of second derivative on age prediction Predicted vs true age for validation set with a) raw data and b) Savitzky-Golay (window=9, second derivative) using PLS. The upper and lower ends of the centre box indicate the 75th and 25th percentiles. The line inside the box indicates the median, and the whiskers show the maximum and minimum no further than 1.5 Interquartile Range (IQR). Letters show groups with statistically different means ($p < 0.05$) by ANOVA with Tukey's multiple comparisons

Our PLS model was able to predict the age of laboratory reared female mosquitoes from IRSSlab using the raw data and pre-processed data (R^2 = 0.77, RMSE = 1.89, R^2 = 0.78, RMSE = 1.89). There was a similar tendency for the model to overestimate the age of young mosquitoes and underestimate the age of older mosquitoes (Table 3.5). Samples were differentiated into five age

Table 3.4: Mean prediction age in *An. gambiae* females using the raw data and Savitzky-Golay pre-processing

Pre-processing	Actual Age	Samples	Mean predicted age [95% CI]
RAW	3	31	4.30 [3.92 – 4.68]
	5	42	7.13 [6.64 – 7.63]
	7	42	7.21 [6.62 - 7.80]
	9	37	9.09 [8.60 – 9.58]
	12	45	10.65 [10.20 – 11.10]
	15	35	12.80 [12.25 – 13.35]
Saviztky Golay	3	31	3.59 [3.07 – 4.11]
	5	42	6.56 [6 – 7.13]
	7	42	7.20 [6.82 – 7.57]
	9	37	9.20 [8.65 – 9.74]
	12	45	10.98 [10.41 – 11.54]
	15	35	14.13 [13.54 – 14.71]

groups as seen with samples of the same origin (Fig 3.8a) and after pre-processing, the samples were differentiated into 6 age groups (Fig 3.8b). However, the model was unable to predict ages from semi-field female mosquitoes. The model overestimates the age of 1, 4, 7 days old by more than 6 days. The mean predicted age for 10 days old was 4.89 days old and for 15 days old was 1.56 (R^2 = -1.84, RMSE = 7.99) Pre-processing did not increase the predictive power of the model, instead the mean predicted age for all age groups was from 26 to 40 days old (R^2 = 33.45, RMSE = 27.86, Table 3.5). A summary of the models and their performance in each data set can be seen in Table 3.6. These results suggest PLS-DA and PLS as a viable model to predict age and species in laboratory reared mosquitoes; with reduced accuracy on samples from other laboratories, but it cannot predict laboratory colonies reared in semi-field conditions.

Here, I presented a baseline performance of the typical chemometrics approach used for species and age prediction using MIRS which has not been addressed before. This approach achieved a similar performance compared to other machine learning models in species prediction and chronological age prediction used with MIRS [17]. Pre-processing increased model accuracy in species and age, which has not been properly evaluated in MIRS studies using *Anopheles*. Poor model generalisation was not fixed by pre-processing, and it is not caused by the model architecture, but because data sets between different labs and rear conditions are fairly different.

Table 3.5: Mean predicted age of laboratory reared *An. gambiae* females (IRSSlab) and field (IRSSfield) used as independent validation sets

IRSSlab			
		Raw	**Savitzky-Golay**
Real Age	**# samples**	**Mean predicted age [CI 95%]**	**Mean predicted age [CI 95%]**
3	31	4.30 [3.92 - 4.68]	3.59 [3.07 - 4.11]
5	42	7.13 [6.64 - 7.63]	6.56 [6.00 - 7.13
7	42	7.21 [6.62 - 7.80]	7.20 [6.82 - 7.57
9	37	9.09 [8.60 - 9.58]	9.20 [8.65 - 9.74]
12	45	10.65 [10.20 - 11.10]	10.98 [10.41 - 11.54
15	35	12.80 [12.25 - 13.35]	14.13 [13.54 - 14.71]
IRSSfield			
1	53	6.96 [6.42 - 7.49]	29.02 [26.43 - 31.61]
4	60	8.84 [8.12-9.55]	33.21 [29.79 - 36.63]
7	55	10.99 [10.02-11.97]	23.32 [20.18 - 26.46
10	48	4.89 [4.16-5.62]	40.55 [36.02 - 45.09]
15	48	1.56 [0.24-2.89]	26.81 [22.64 - 30.98]

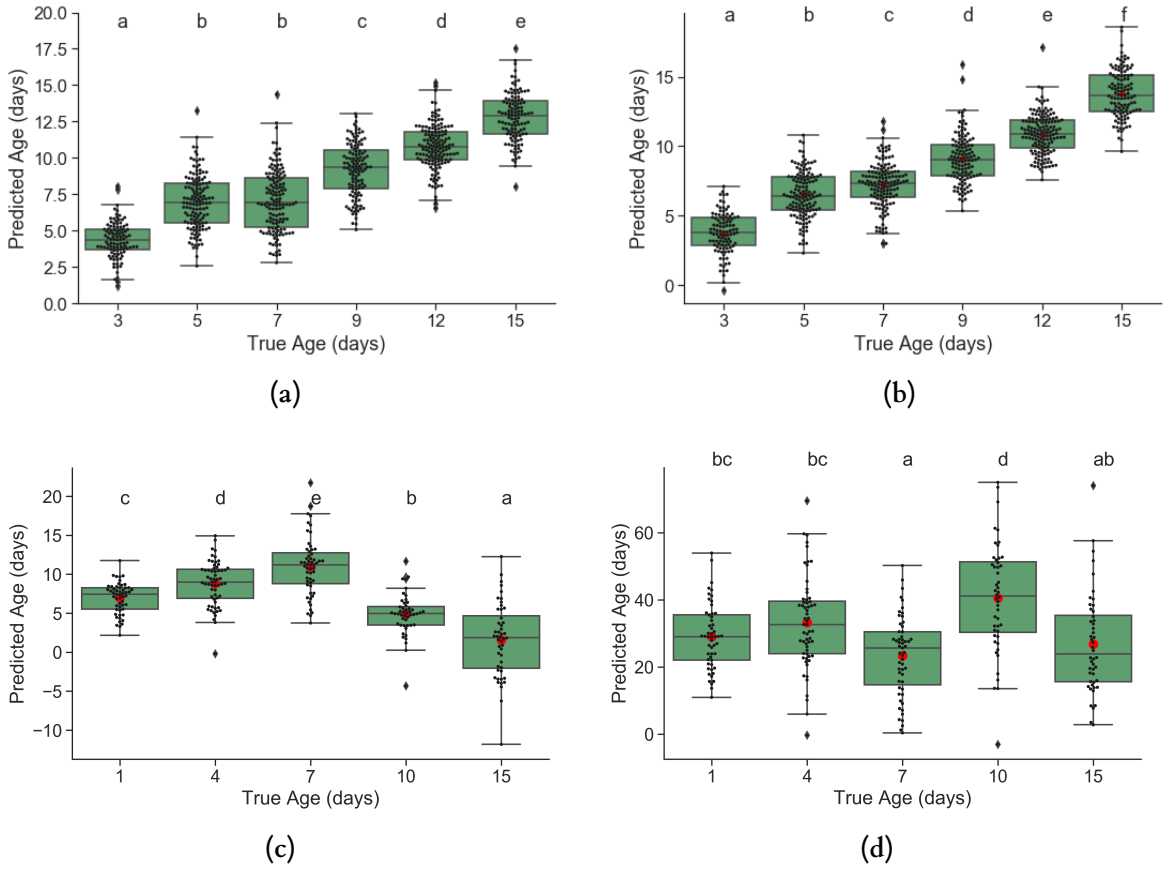


Figure 3.8: Effect of second derivative in age prediction with independent validation sets Comparison of predicted vs actual age using a subset of Glasgowlab data set as independent validation set with a) raw data and b) after applying Savitzky-Golay pre-processing. Comparison of predicted vs actual age using a IRSSfield data set as independent validation set with c) raw data and b) after applying Savitzky-Golay pre-processing

Table 3.6: Mean predicted age of *An. gambiae* mosquitoes using IRSSlab and IRSSfield datasets as independent validation sets

# Samples	LV	Pre-processing	RMSE Cal	R^2 Cal	RMSE CV	R^2 CV	RMSE V	RMSE v IRSSlab	R^2 IRSSlab	RMSE V IRSSfield	R^2 IRSSfield
772	15	RAW	2.24	0.68	2.47	0.60	2.16	1.89	0.77	7.99	-1.84
772	14	Savitzky Golay	1.88	0.77	2.23	0.68	1.91	1.88	0.78	27.86	-33.45
772	16	SNV	2.14	0.70	2.37	0.64	2.10	n/a	n/a	n/a	n/a
772	15	MSC	2.16	0.70	2.39	0.63	2.14	n/a	n/a	n/a	n/a
772	17	RNV	2.15	0.70	2.41	0.62	2.22	n/a	n/a	n/a	n/a

3.3.4 The influence of scatter correction algorithms on machine learning models accuracy for species prediction

Scatter correction algorithms were applied to the data to reduce the effect of scattering between the samples. These were multiplicative scatter effect, standard normal variate robust normal variate and Savitzky-Golay. The results in figure 3.9 suggest pre-processing can increase model accuracy compared to the raw data. All models showed an increase in accuracy with NB benefited the greatest with an increment from 0.65 to 0.8 with SNV, MSC and up to 0.81 with Savitzky-Golay. The reason is that pre-processing removes any unwanted variability and enhanced the most important features for our classification problem. However, this increase in accuracy depended on the model architecture. LR and SGD which are linear models showed a decreased in accuracy to the same level as a random model (accuracy = 0.5) when Savitzky-Golay was used. The extreme changes to the spectrum (smoothing and second derivative) caused by Savitzky-Golay might erase all the differences between species, making it extremely difficult for those models to separate the two classes. Interestingly, nonlinear models (NB, RF, CART) benefited more from the extreme pre-processing. The same phenomenon was seen when models were tested on hold-out data set (Fig. 3.10). Here, scatter correction not only increased accuracy but also improved group classification across all models, suggesting that the elimination of the noise caused by scattering enhance features which are important for species prediction.

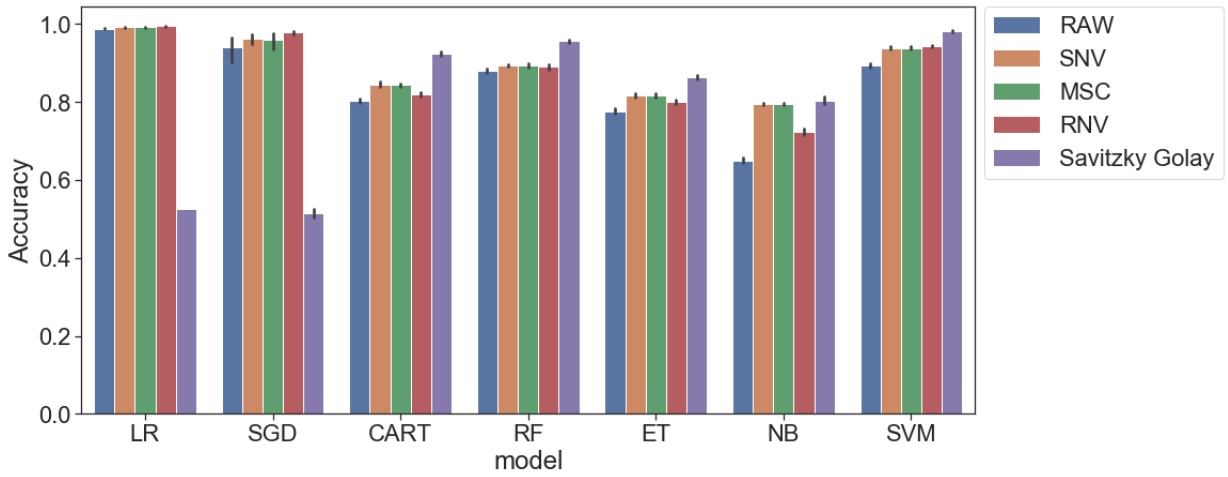


Figure 3.9: Accuracy of different machine learning models when different scatter corrections were applied: No pre-processing (RAW), multiplicative scatter correction (MSC), standard normal variate (SNV), robust normal variate (RNV) and Savitzky-Golay. Models were trained using Glasgowlab dataset and tested using 10-fold cross-validation.

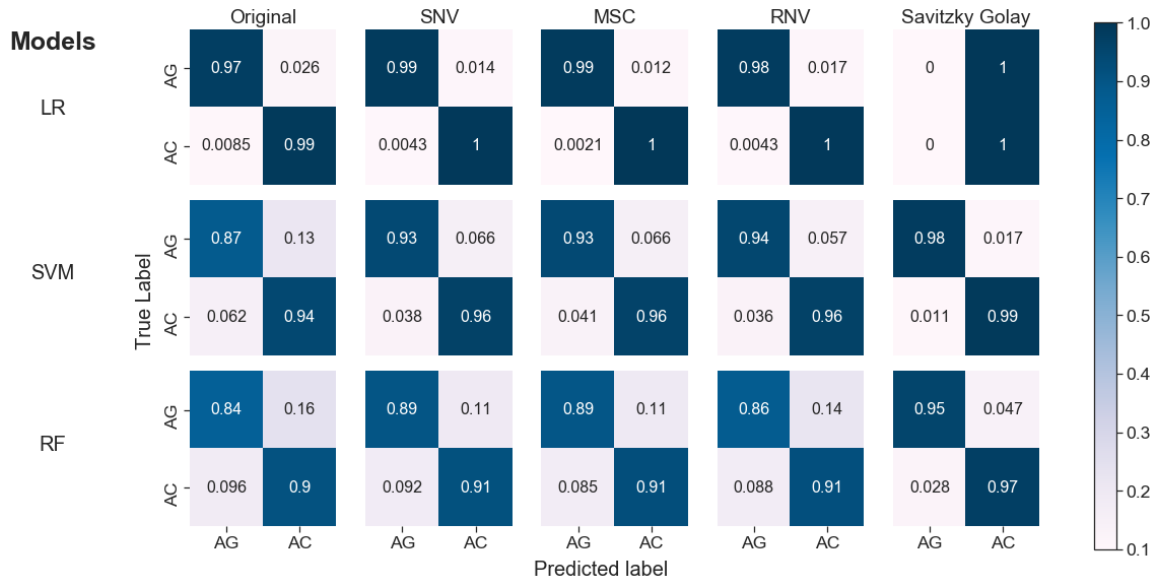


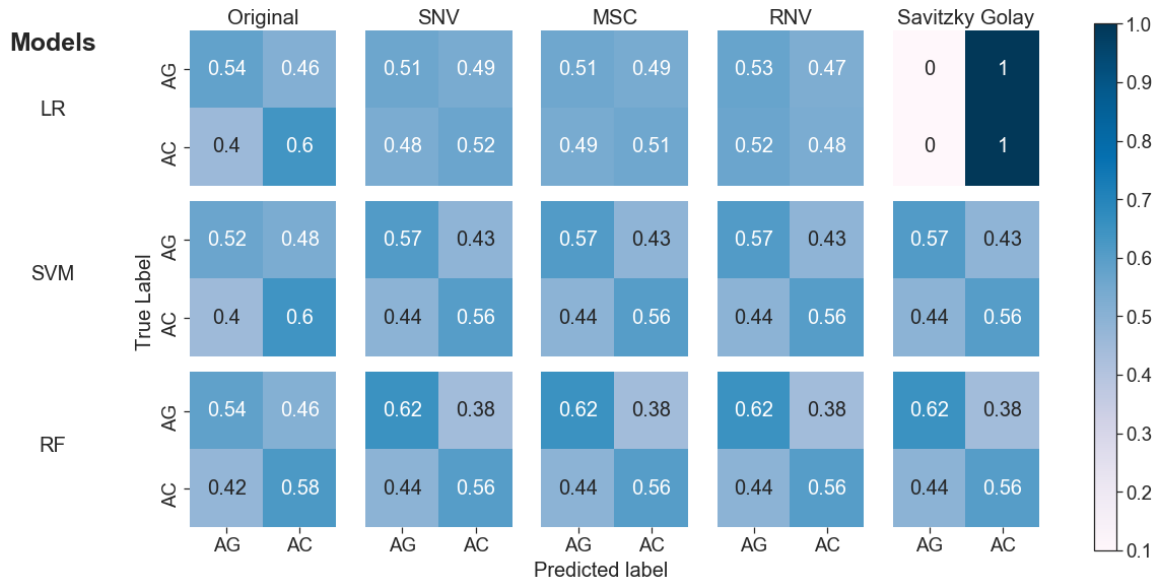
Figure 3.10: Normalised confusion matrix for final models trained in Glasgowlab dataset and tested on a hold out set. Each row represents instances of true class, while each column represents instances of predicted class. AC: *An. coluzzii*. AG: *An. gambiae*

Effect of scatter correction algorithms on model accuracy with independent validation sets for species prediction

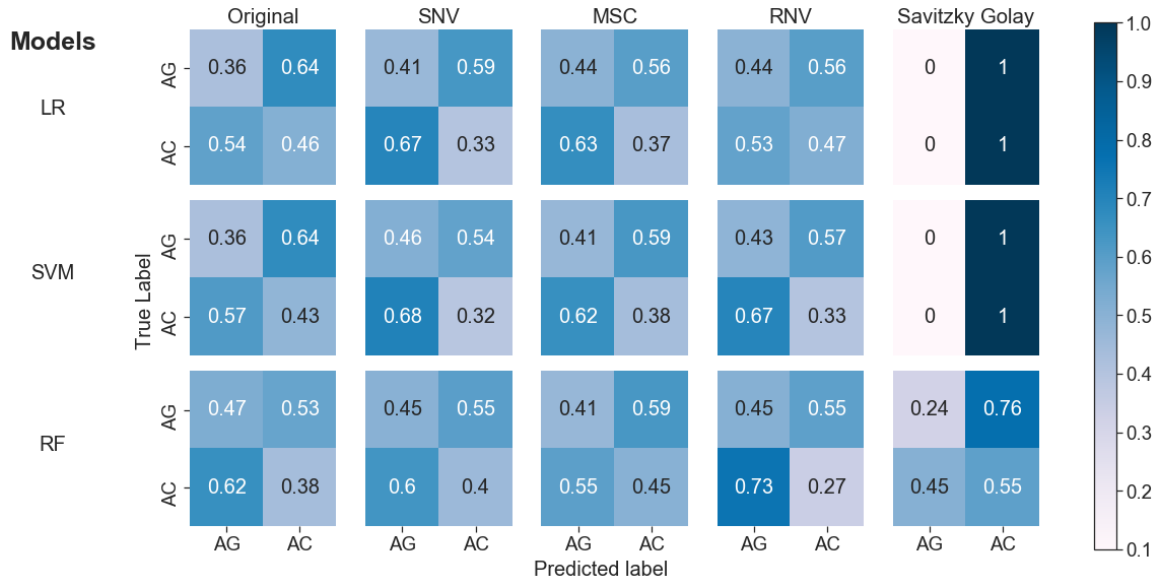
Quantification of the effect of scatter correction on classification in species prediction was assessed using two independent sets: IRSSlab and IRSSfield data sets. The classification accuracy drops below 0.6 on all models (Table 3.7). LR did not benefit from pre-processing, reaching the same accuracies as a random model when Savitzky-Golay was applied. The same phenomenon was observed using SGD. Pre-processing slightly increased classification accuracy in the RF model compared with the un-processed spectra. Confusion matrices show SVM provided slightly better classification in species prediction than the rest of all pre-processing algorithms (Fig. 3.11a). Validation on IRSSfield samples showed the worst prediction accuracy (Table 3.7). Confusion matrices show how species are misclassified by all the models with un-processed and pre-processed data (Fig. 3.11b.) The reason for the low model performance is that while pre-processing eliminates differences between samples to maximise class separation, there are inherent differences between the data sets which are being maximised by pre-processing algorithms. This makes model generalisation extremely difficult.

Table 3.7: Species prediction accuracy of LR, SVM and RF using independent test sets IRSSlab and IRSSfield

Preprocessing	Model	IRSSlab	IRSSfield
RAW	LR	0.57	0.41
	SVM	0.56	0.39
	RF	0.56	0.43
SNV	LR	0.52	0.37
	SVM	0.57	0.39
	RF	0.59	0.42
MSC	LR	0.51	0.41
	SVM	0.57	0.39
	RF	0.58	0.43
RNV	LR	0.50	0.45
	SVM	0.57	0.38
	RF	0.59	0.36
Savitzky Golay	LR	0.50	0.49
	SVM	0.50	0.49
	RF	0.56	0.39



(a)



(b)

Figure 3.11: Normalized confusion matrix for final model on validation sets: **a)** IRSSlab and **b)** IRSSfield with different pre-processing algorithms. Each row represents instances of true class, while each column represents instances of predicted class. AC: *An. coluzzii*. AG: *An. gambiae*

3.3.5 Analysis of different spectral windows for species prediction

In order to select the most informative region of the spectra for future applications (such as QCLs), I divided the spectra into 3 regions (Fig. 3.12). Each window was used as input to three different ML models trained. Glasgowlab dataset was used as training set and the effect of the use of different windows on prediction accuracy was quantified by the performance of each model validated with the three data sets: Glasgowlab (hold out set), IRSSlab and IRSSfield. LR had the highest

mean accuracy when the whole range (1800 to 950 cm^{-1}) and X1 region were used (0.98 and 0.98, Fig 3.13), when validation was done with samples from the same origin. Prediction accuracy decreased when the window moved into lower wavenumber values (accuracy X2 = 0.91, X3 = 0.85), however, it never dropped below 0.80. SVM and SGD accuracy was higher when using X1 (0.87, 0.94) compared to the entire region (0.86, 0.92). SMV classification was the lowest of the models (0.67) when using the X3 window. These changes in accuracy when using different regions of the mid-infrared pointed to the X1 window as the one with more information for species prediction.

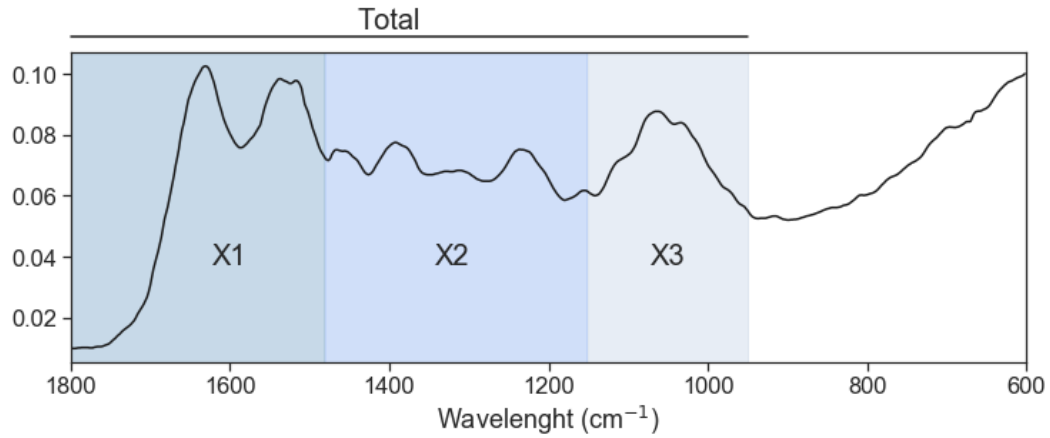


Figure 3.12: Division of the mid infrared spectrum into 3 different windows from 1800 - 1500 cm^{-1} (X1), 1500 - 1250 cm^{-1} (X2) and 1250 - 950 cm^{-1} (X3).

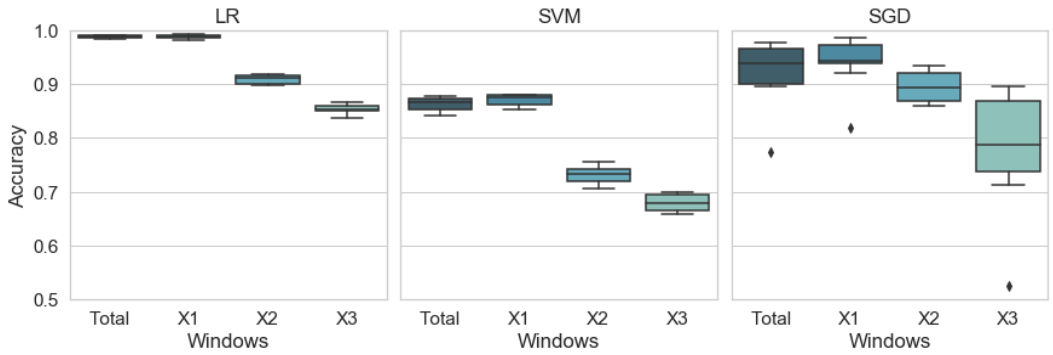


Figure 3.13: Boxplot of model accuracy in cross-validation for LR, SVM and SGD when using the whole mid infrared spectrum (Total) and each of the spectral windows (X1, X2, X3)

X1 window has the Amid I and II bands and ester carbonyl, which are related to chitin, fatty acids, wax and proteins [17, 222]. The higher accuracy when using this window suggest the majority of differences between the two species are due to changes in protein, chitin and wax content. These bands have been shown to be related to age in *An. gambiae* [199] and *Ae. aegypti* [222]. Consequently, this window was further divided into three smaller sections (X4, X5 and X6) to examine if there is a specific region inside the Amid I and II bands which is responsible for the high accuracy reported here. The X5 window, which contains the Amide I band and C=O band showed the highest accuracy for species prediction of 0.98. *Anopheles gambiae* can be

discriminated with 0.96 and *An. coluzzii* with 0.99 (Fig. 3.14). The region X4 with Amide I and II bands reached 0.94 followed by X6 with 0.88 using LR 3.15. The prediction coefficients of the X5 window show that most weighted wavelengths are located in the 1700 - 1775 cm^{-1}region, where the Amide I and C=O band are located. The ester carbonyl band is related to fatty acids, suggesting that the difference in fatty acids and chitin content between the two species allows them to be differentiated with great accuracy.

Results of the evaluation of each of the window with IRSSlab and IRSSfield data sets as validation sets shows the same trend of decreasing accuracy below 0.60 and 0.50 continued (Table 3.8). Interestingly, some models showed similar accuracies when validated with IRSSlab and IRSSfield. For example, SVM showed consistency accuracy across both validation data sets even when small windows (X4, X5, X6) are used, the same as with SGD. SGD shows an accuracy of 0.58 when using X4 window. For comparison, PLS-DA accuracy across all windows were added. As with the other algorithms, the X4 and X5 regions showed the highest accuracy, similar when using the entire spectrum (Appendix B, Fig. B.6). By removing regions with less information, it is easier for the model to separate the two species. This might suggest that using small wavelength ranges decreases the variation between datasets decreases and retain some species-specific which can be identified by classification algorithms.

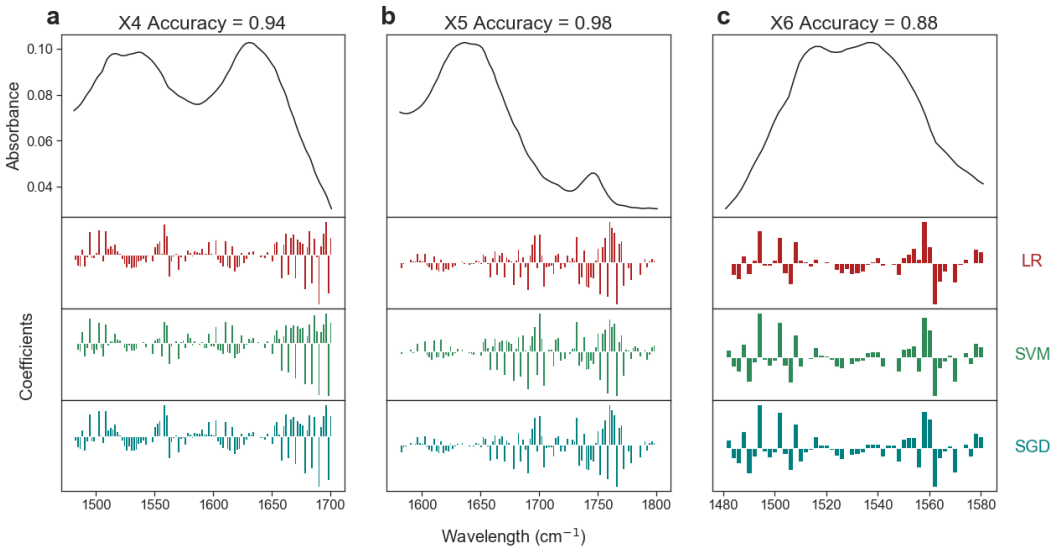


Figure 3.15: Accuracy and prediction coefficients of three models logistic regression (LR), support vector machine (SVM) and stochastic gradient descent (SGD) for each window: **a)** X4 (1700-1500 cm^{-1}), **b)** X5 (1800-1600 cm^{-1}) and **c)** X6 (1580-1480 cm^{-1}).

Figure 3.14: Accuracy on hold-out data set. Normalised confusion matrix for each model on hold-out data set. Model tested are logistic regression (LR), support vector machine (SVM) and stochastic gradient descent (SGD). Each row represents instances of true class, while each column represents instances of predicted class. AC: *An. coluzzii*. AG: *An. gambiae*

Table 3.8: Accuracy of species prediction of LR for each spectral window in cross-validation (CV), validation (Val) and validation with independent data sets (IRSSlab, IRSSfield)

Window	Model	CV	Val	IRSSlab	IRSSfield
X1	LR	0.99 (±0.01)	0.98	0.51	0.37
	SVM	0.87 (±0.01)	0.88	0.54	0.54
	SGD	0.94 (±0.06)	0.89	0.44	0.44
	PLS-DA	0.85 (±0.02)	0.84	0.58	0.45
X2	LR	0.91 (±0.01)	0.90	0.58	0.36
	SVM	0.73 (±0.01)	0.73	0.52	0.51
	SGD	0.89 (±0.06)	0.86	0.56	0.56
	PLS-DA	0.83 (±0.02)	0.82	0.56	0.43
X3	LR	0.85 (±0.01)	0.86	0.55	0.30

Table3.8 *(Continued)*

Window	Model	CV	Val	IRSSlab	IRSSfield
	SVM	0.68 (±0.01)	0.68	0.46	0.45
	SGD	0.78 (±0.06)	0.78	0.55	0.55
	PLS-DA	0.79 (±0.02)	0.80	0.61	0.30
X4	LR	0.94 (±0.01)	0.94	0.57	0.38
	SVM	0.82 (±0.01)	0.83	0.55	0.55
	SGD	0.92 (±0.06)	0.97	0.58	0.58
	PLS-DA	0.87 (±0.01)	0.87	0.59	0.49
X5	LR	0.99 (±0.01)	0.98	0.51	0.35
	SVM	0.89 (±0.01)	0.88	0.56	0.56
	SGD	0.95 (±0.06)	0.96	0.47	0.47
	PLS-DA	0.85 (±0.01)	0.84	0.59	0.49
X6	LR	0.87 (±0.01)	0.88	0.53	0.43
	SVM	0.73 (±0.01)	0.74	0.52	0.51
	SGD	0.88 (±0.06)	0.93	0.56	0.56
	PLS-DA	0.86 (±0.02)	0.86	0.57	0.49

Evaluation of different spectral windows on IRSSlab data set for species prediction

To assess whether the trend of decreasing accuracy in species prediction in different spectral windows can be seen in another data set, I used the IRSSlab data set exclusively for analysis. As expected, the highest accuracy was achieved when using the complete spectrum (windows X1+X2+X3) regarding the model used in cross-validation (Fig. 3.16) or hold out set (Fig. 3.17). However, there was no accuracy variation between the different windows. X2 window showed slightly higher accuracy compared to the other windows when using LR, while there was no difference in accuracy between windows with SVM and SGD. However, when X1 and X2 were combined, accuracy increased almost at levels when using the complete spectrum X1+X2 (Fig. 3.16 and Fig. 3.17). The reason is that the difference between species is most likely located in the Amid I and II bands (X1 region) and also in bands related to chitin/wax related (X2). This

was further confirmed when examining the prediction coefficients (Fig. 3.18). The model coefficients showed the relative importance of wavenumber values on the decision (species prediction). IRSSlab most weighted prediction coefficients (most important) are distributed across the 1600 to 1400 cm^{-1} region (Windows X1 and X2). This distribution is different when comparing to prediction coefficients in Glasgowlab where its most weighted wavelengths are located at 1800 cm^{-1} to 1700 cm^{-1} (Window X5) region. The differences between species in the IRSSlab data set are exclusively due to changes in proteins and chitin (Amid I, O=C-N, C-CH_3) and less due to C=O and C-O bands which are chitin and wax only. On the other hand, in the Glasgowlab data set, differences between species are caused by changes in fatty acids and proteins. These results show that high accuracy for species prediction (> 80%) can be achieved by only using the region from 1800 to 1250 cm^{-1} ($\approx$ 650 cm^{-1}). Moreover, the differences in important features between the two data sets in species prediction are the cause of poor model generalisation.

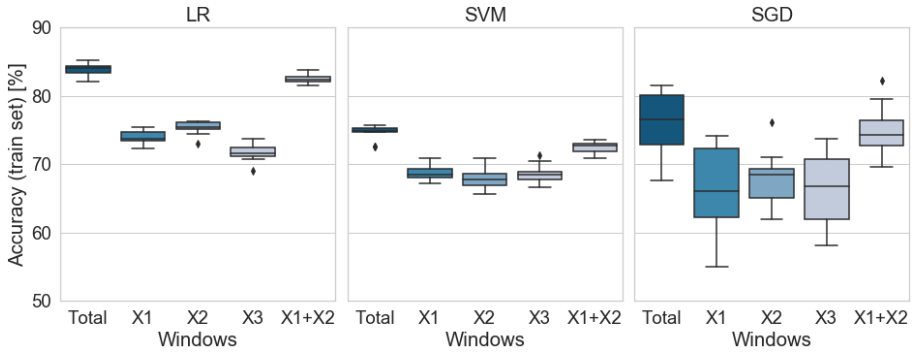


Figure 3.16: A) Baseline accuracy. Classification accuracy after 10-fold cross-validation using logistic regression (LR), support vector machines (SVM) and stochastic gradient descent (SGD) with different mid-infrared regions: total spectra (1800 - 950 cm^{-1}), X1 (1800 - 1500 cm^{-1}), X2 (1500 - 1250 cm^{-1}), X3 (1250 - 950 cm^{-1}) and the sum of X1 and X2 (1800 - 1250 cm^{-1}).

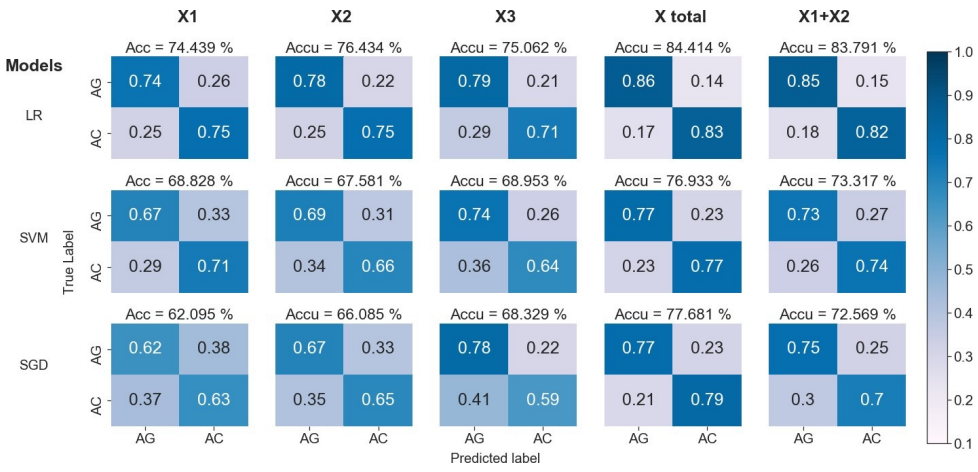


Figure 3.17: A) Validation on hold out set. Normalised confusion matrix for final models on validation tested on hold out set. Each row represents instances of true class, while each column represents instances of predicted class.

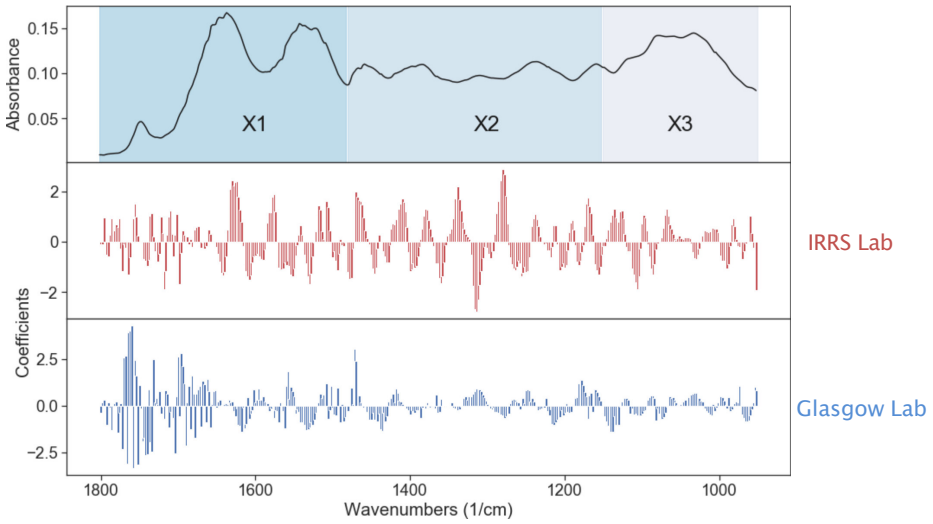


Figure 3.18: B) Variable importance. Comparison of prediction coefficients between IRSSlab (red) and Glasgowlab (blue) data sets for each wavenumber. High values (positive or negative) suggest a wavenumber is more important on the model decision. The most import features are located in the X1 region for Glasgowlab data set compared to IRSSlab dataset where the most important features are located across x1 and X2 regions.

3.3.6 Analysis of spectra between datasets using Principal Component Analysis

Analysis of PCA found two principal components which explained 77.82%, 16.83% of the variance observed in all data sets combined (Fig. 3.19a). PCA scores plot shows how IRSSfield clustered towards PC1 negative values (Fig. 3.19b). Differences between the laboratory samples and semi-field samples are more evident when comparing scores of PC1 (Fig. 3.19c). Based on PC1 loadings, wavelengths in the Amide I and II region and the absorption band at 1000 cm^{-1}) have negative loadings, while C = O band (1750 cm^{-1}) and 1500 to 1000 cm^{-1} have positive loadings (Fig. 3.19c). Glasgowlab have higher absorption intensity in the two regions with positive loadings and those intensities decrease in IRSSlab as seen in the downright trending of PC1 score values. These differences are also seen when comparing the mean spectra of the three databases (Appendix B, Fig. B.9). Overall, the differences in absorption bands related to proteins, waxes and chitin present in samples from IRSSfield may explain the difficulty of predicting age and species when models are calibrated with laboratory-reared samples and tested on semi-field reared samples. The reason is that the model learns specie-specific differences from the laboratory-reared samples which are located in a very limited region of the spectrum, while the differences in the samples from other laboratory and semi-field reared samples are located in other regions, making it very difficult for the model to predict species.

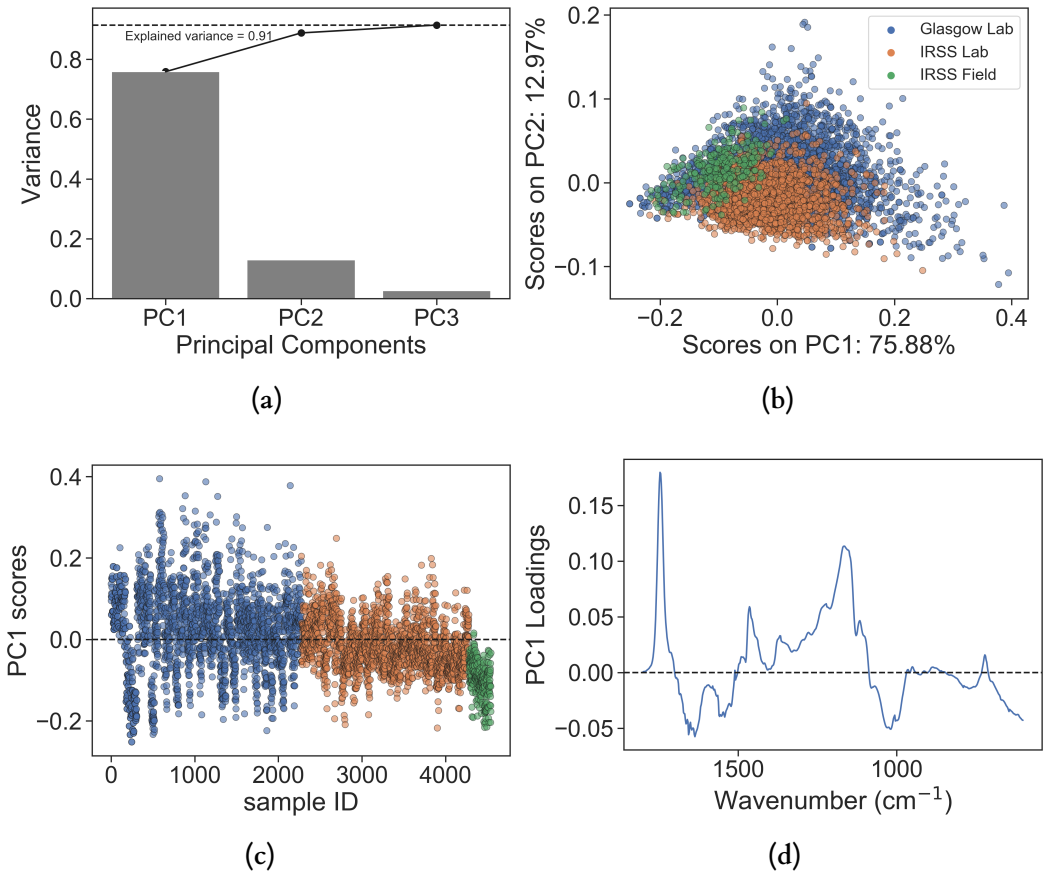


Figure 3.19: a) Explained variance (barplot) and cumulative explained variance (line dot) of each PC1, PC2 and PC3. b) PC1 versus PC2 scores plot mean centred spectra of samples from Glasgowlab (blue) and IRRSlab (orange) and IRSSfield (green). There is a separation between IRSSfield and IRSSlab and Glasgowlab. **b)** Scatter plot of PC1 scores by sample. **c)** PC1 loading versus wavenumber values.

3.4 Discussion

Here, I have presented a thorough analysis of ATR-FTIR spectral data from laboratory and field mosquitoes. I assessed the potential of the most common approach, the use of partial least squares, against less common machine learning algorithms for species and age prediction. Furthermore, I analysed different scatter correction algorithms and their effect on model performance in prediction of *Anopheles* mosquito species classification and age grading. I also quantified the effect of using different spectral regions for analysis and the differences between samples from two different laboratories and field samples.

PLS has been used mainly in three other studies on *Aedes* mosquitoes for age grading, *Wolbachia* infections [222] and species identification [224]. PLS shows high accuracy in binary classifications, and our model reported similar results for species prediction (overall accuracy = 0.87). These values are comparable with the ones obtained by Gonzales et al. [17] for *An. gambiae* and *An. arabiensis* with accuracy of 0.82 and 0.84 respectively. There are no studies for *An. coluzzii* using MIRS, but the classification accuracy of our model is above 0.9. Although, PLS has not been used in *Anopheles* and MIRS, PLS has been used consistently through most NIRS studies and our results are comparable to them [18, 19]. Therefore, PLS or ML algorithms for species classification show similar performance when samples from the same origin are used for training and validation.

PLS was able to discriminate between mosquitoes aged from 1 to 17 days old with overall moderate accuracy (R^2=0.68). This is higher compared to Krajacich, et al. (R^2CV=0.39) [214] when comparing the same algorithm and validation with samples from the same origin. The R^2 values here are only surpassed by an ensemble PLS model (R^2CV=0.74). Moreover, I was able to discriminate 5 out of the 6 age categories used (without pre-processing), and 6 out of 6 (with pre-processing); whereas Krajacich, et al. just discriminate 4. This was also true in age prediction with *Aedes* [225]. However, there was a consistent tendency for over-estimation of the age of young mosquitoes, and underestimation of older mosquitoes. Similar to Gonzales, et al. [17], very young mosquitoes can be discriminated but in contrast to that study, our 15 days old mosquitoes were underestimated. Mean predicted age was 14.13 [95% CI 13.54 – 14.71]. This over/underestimation is caused by the use of PLS, which is a linear model and chronological age is not linear. Overall, our results for species prediction are in line with previous studies in NIRS and MIRS. For age prediction, the results are slightly better, considering the use of a simpler PLS model. This might be due to the amount of information (absorption bands) in the mid-infrared spectra compare to near infrared. Near infrared spectra consist of overtones and combination of bands from the mid infrared region. These bands overlap to each other a not well resolved compared to PLS region. This might mask some information, resulting in lower model performance. I chose a wider gap between ages since, probably, there is not a significant change of the chemical profiles between

close age groups. The issue might be trying to predict chronological age, which is how humans perceive time, of a biological system, without understanding how it relates with physiological age and changes in the spectrum. When comparing true age and predicted age (Fig. 3.5, we can differentiate three groups: 1 to 4, 5 to 9 and 10 to 17. Ages with 1 day apart are very similar, suggesting that the spectra do not change within those intervals. Studies that tried to predict age with a higher resolution have failed in MIRS [17] and NIRS [214, 220] and it is common to wider the gap between groups with intervals of 2, 3 or 4 days, not only in MIRS/NIRS but also in Raman [223] and MALDI-TOF mass spectrometry [211].

One of the problems with NIRS is the decrease in model prediction when tested with samples from other laboratories or field samples [214]. However, a recent study by Milali, M. et al. 2020 [218] showed ANN models with autoencoders can predict parity status from other cohorts which were not used in model training. My results suggest that MIRS suffers from a poor prediction with datasets from other origins and rear conditions. Although, high accuracy of age prediction between different laboratory samples was achieved and to a lesser extent for species prediction, models were incapable of predicting species and ages from laboratory samples reared in semi-field conditions, independently of the model used. I expected at least good classification power for species when using the IRSSlab dataset as these mosquito samples are reared in similar laboratory conditions in terms of temperature (27 °C) [17, 310] but it was not the case. The use of different strains between laboratories might be playing a role in the decrease of species prediction. High coefficient values for Glasgowlab were located mainly at the 1750 cm^{-1} region for species (Fig. 3.18) while IRSSLab coefficients were distributed more evenly between 1800 and 1200 cm^{-1} region. Although, genetic (different strains) might be affecting the prediction power, rearing conditions seem to have the most influence, since the lowest accuracy was obtained when trying to predict species and age from the IRSSfield. Temperature has an effect on CHCs profiles in *Drosophila* [311, 312]. Higher temperature exposure is related to higher concentration of longer chain CHCs, as well as short-term changes in temperature and other selection pressures (climatic conditions, physiological constraints) [313]. This is also seen in *Anopheles*, where humidity and photoperiod changes CHCs concentration [314]. On top of that, colonies from different laboratories might be highly genetic divergent [315]. These changes in cuticle composition are reflected in the spectra with differences in the amid I and amid II bands and notably around 1000 cm^{-1} region (Appendix B, Fig. B.8 and Fig. B.9). These changes might mask the differences between species and age. Overall, the lower variability within individuals from the same laboratory specimens compared to counterparts from other lab or field populations, plus temperature differences in rearing and holding conditions between laboratories, affect drastically model performance in age and species prediction. Therefore, model training needs to include a subset of semi-field and field derived mosquitoes in order to improve model generalisation as shown in recent studies [199]

Scatter corrections highlighted that the Savitzky-Golay correction had a detrimental effect in

LR and SGD models even in cross-validation, but led to an improvement when using with PLS. However, the effect of Savitzky-Golay was detrimental when validated with independent data sets. Decreased in model performance due to Savitzky-Golay has been seen in crickets [316]. Pre-processing tends to increase the differences between groups from the same sample batch, however, extreme pre-processing eliminates the signal differences between groups making it difficult for the model to find any difference, therefore decreasing its performance

Research in MIRS for mosquito surveillance is expanding, and is characterised by a move towards more complicated algorithms such as RF or Artificial Neural Networks (ANN). However, there is evidence that PLS can reach similar prediction performance compared to ANN [119,317] or RF and SVM [318] but when non-linear relationships appear, ANN can be more useful [119]. These results also support this interpretation; PLS reached similar results when compared to other machine learning models for species prediction, but for prediction of chronological age, non-linear models (ANN) might perform better as suggested on recently NIRS studies [218,219]. Therefore, I suggest the use of PLS as first method for species prediction and if the prediction performance is not satisfactory, move to more complicated models (Logistic regression for example). For age prediction, if dealing with chronological age or gonothropic cycles, ANN might be more suitable as a starting point.

Further work is needed to assess how temperature affects spectral data. One way of doing this is to reared mosquitoes at different temperatures and humidity while other conditions constant remain (food, body size, etc.) and assess which bands change and how they change (peak height and width). Doing this will give us insight on what specific compounds in the cuticle are changing according to temperature, and it might be possible to find a relationship between temperature, changes in the spectrum, age, and species. Currently, there is limited information on how MIRS spectral data vary between laboratory colonies reared in different conditions; in response to temperature or other local environmental conditions. However, my results differ from those reported in a previous study with NIRS where no fundamental differences in spectra between laboratory reared and wild mosquitoes were found using k-means [319]. However, two studies have produced similar findings that age models derived from laboratory data do not work when applied to field samples or field samples reared in semi-field conditions [214, 217]. The first of these was assessing age grading in *Anopheles* [214], and suggested genetic variability and a lack of control over variables such as food were responsible for this phenomenon. The second study hypothesised this was due to differences in water content between laboratory and field derived mosquitoes [217]. Water content can be a source of variation in NIRS, as usually samples are not dried before spectra measurements. Differences in water content between ages groups in insects has been reported using NIRS, where younger weevils have higher water content compared to old ones [320]. Moreover, OH band related to water is a major band used in the detection of wheat pests [274,321]. Therefore, water might play a role in age and species prediction in mosquitoes. However, the potentially confounding impact of water content was minimised in our samples

because they were desiccated before analysis. My results support the hypothesis of fundamental differences between laboratory samples reared in laboratory and semi-field conditions, and even between samples from different laboratories. Therefore, due to the data mismatch, it is imperative the use of semi-field samples or samples from different laboratories in the model calibration.

Finally, I have shown the viability of using much narrower spectral regions of around 300 cm^{-1} for species prediction. Although the validation accuracy decreased as the window moved to lower wavenumber values, it never went below 70%; and remained similar across windows for the independent sets. One thing to highlight is the lower accuracy achieved by LR compared to SVM and SGD, which remained consistent. Except when using windows X3. I hypothesised that if model training includes more diverse samples from different laboratories and semi-field conditions, the accuracy for each window should increase as well. This will open up opportunities to implement new technologies, such as Quantum Cascade Lasers, which is emerging as an alternative to more traditional globar FTIR in the field of biomedical science for tissue classification. These lasers have a narrower but customise wavelength ranges. Therefore, a Quantum Cascade Laser-based spectrometer can be built according to a specific spectral window of interest for more efficient data acquisition.

3.5 Conclusion

This study showed how traditional chemometrics can be applied to ATR-derived spectra data for species and age prediction in the malaria vector *An. gambiae*. It also showed how the issue of lower prediction power when using training and validation data sets come from different origins is caused by the mismatch (data for training is not representative of the data that will be used when the model is deployed) and not by the model used. Moreover, it shows the benefit of exploring narrower wavelengths and its impact in model prediction and its potential for a more efficient way of data extraction and open the opportunity to develop custom spectrometer using next generation mid-infrared light sources. In summary: i) PLS and PLS-DA can be used as models for species and age prediction when using laboratory samples. ii) Scatter corrections on machine learning models can increase model performance but Savitzky-Golay should use with caution with models such as LR. iii) The uniformity of laboratory samples, in addition to differences in rearing conditions, have an impact in the absorption bands mainly in the amide I, II region and around 1000 cm^{-1}. These changes are in agreement on how CHCs changes according to temperature, physiological factors reported in previous studies. iv) Narrower spectral regions of 300 cm^{-1} showed potential for species prediction, especially the Amide I and II region. Therefore, there is still room for exploration. The question of how temperature affects the spectra data needs to be addressed, in order to be considered when dealing with field samples. New techniques such as transfer learning [310] can fix the problem of data mismatch by adding a small

percentage of wild caught or semi-field samples to the training set to improve model generalisation. Future work should focus on how different spectral windows can be applied according to different prediction problems.These will help us to move spectroscopy as a surveillance tool forward into better and faster devices.

Chapter 4

Portable, Fast-swept External Cavity Quantum Cascade Laser system for Spectroscopy

4.1 Introduction

Infrared spectroscopy is one of the most important and versatile techniques for chemical analysis [25]. It characterises the chemical composition of a sample by measuring the absorption of light by exciting its chemical bonds [322]. This is possible thanks to the capacity of functional groups to absorb infrared radiation due to changes in the dipole moment during vibration [323]. Infrared spectroscopy therefore probes these vibrational modes [20] and, importantly, a given functional group absorbs light at the same frequency independent of the molecule it is attached to [33]. Every different functional group absorbs at a different frequency, so an absorption spectrum of a biological sample contains a set of the unique absorption signatures indicating the presence of proteins, carbohydrates, lipids and nucleic acids [24, 107].

Fourier transform infrared spectrometers (FTIR) are the industry standard for mid infrared spectroscopy across a variety of areas and applications [54]. They traditionally use Globars as a light source that emit broadband radiation, but this extended incoherent source does not allow for compact systems or high resolution microscopy without significant loss of signal-to-noise ratio. Nonetheless, the use of mid-infrared spectroscopy for mosquito surveillance [17, 222, 224, 229] and malaria diagnosis [228, 230] using FTIR have shown promising results in laboratory and field settings. However, the equipment is not without its limitations for this application. The infrared light is emitted with a low optical power (μW range) which can limit its use with high absorbance samples [53]. Moreover, it lacks spatial coherence and requires liquid nitrogen cooled HgCdTe

(MCT) detectors to achieve high signal-to-noise ratio [324,325]. Miniaturisation of FTIR equipment is also challenging due to the presence of a delicate interferometer with a physical mirror scanning system [107, 326]. For example, the resolution of the spectrum is directly determined by the path length difference in the interferometer. For higher resolutions, the path difference needs to be larger. Current portable FTIR equipment offers spectral resolution not better than 4-8 cm^{-1} [327] while FTIR spectrometers based on micro-electromechanical systems (MEMS) have reported resolutions of 66 cm^{-1} [328], which is impractical for our purposes. Other challenges are power consumption, IR sources and infrared detectors [327].

Since Quantum Cascade Lasers (QCL) were demonstrated in 1994 [63], they have already challenged conventional FTIR thanks to the advances in performance at room temperature [73], power output [329] and wall-plug efficiency [330]. They offer compactness [331,332], high optical output power [52], comparable SNR to FTIR when using simple non-cryogenic thermoelectrically-cooled detectors and faster spectral data acquisition [325, 333]. Numerous studies have reported an increased speed and resolution in chemical imaging [102,103,334], improved analysis of amide I and II bands in proteins [53,335], increased path length for aqueous solutions [51] and accurate protein secondary structure monitoring [336, 337] Moreover, QCLs are the core of the development of portable and fast systems beyond biomedical applications such as stand-off chemical sensing [338, 339] and quantification/classification of microplastics [116] which showcases the versatility of these novel laser devices.

As evidence of mid infrared spectroscopy as a reliable tool for mosquito surveillance is increasing, it is imperative to explore alternatives to the FTIR to improve and adapt it to the challenges in the field. We need to build custom, compact, and field ready systems that can be deployed anywhere. QCLs already have the characteristics needed to be the infrared light source for these new systems. Building on current developments in room temperature operating continuous wave QCLs, here is discussed a compact spectrometer using an external Cavity (EC) QCL-based spectrometer in the 9-10.5 µm MIR region aiming for mosquito surveillance. First, I will describe the system design, laser characterisation and system performance. After that, I will show the results of mid-IR spectroscopy measurements of different samples. I shall benchmark the QCL system performance against a commercial FTIR on the same set of samples.

4.2 Overview of the system

The spectroscopy system is designed for fast spectroscopy measurements at 9 - 10.5 µm using an external cavity quantum cascade laser configuration. It was built by me in the Photonics Devices and Systems laboratory at the University of Glasgow. It costs approximately £18000 excluding the computer. It has the following measurements: length x width x height = 300 mm

x 600 mm x 150 mm, and its weight is 22 lbs. The system consists of a quantum cascade laser, a cooling system, an external cavity (diffraction grating, scanning galvanometer) and a photovoltaic HgCdTe (MCT) detector (Fig. 4.1). The system is controlled by software using LabVIEW with in-house developed scripts. All components are mounted on an easy to carry optical baseplate with handles, except for the drive electronics units.

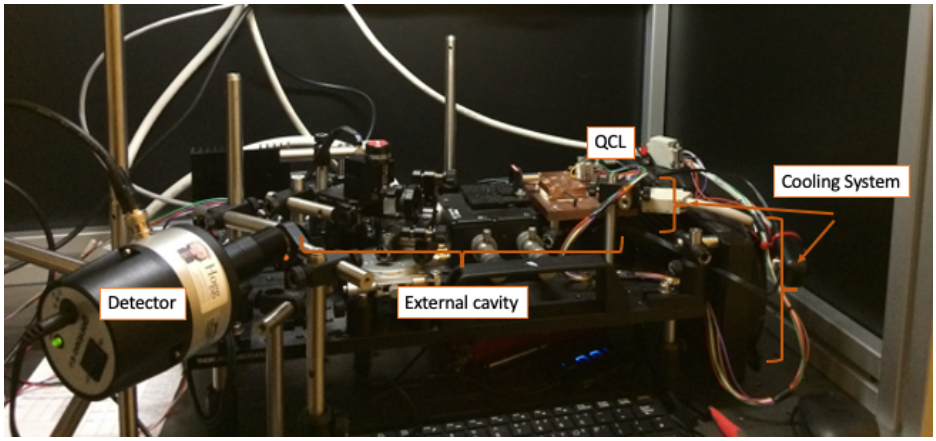


Figure 4.1: Photography of the spectrometer. The optical table houses the main components of the set-up: Laser, external cavity and galvo mirror. Temperature and power sources, detector and sample holder are located externally in this instance. The size here is determined by required flexibility of adaptation in the lab and the size of off-the-shelf mechanics and optical mounts built on a traditional breadboard. It can be significantly miniaturised down to the 'matchbox' level by integration of the mechanical parts onto a micro-bench.

4.2.1 Laser and external cavity configuration

The optical configuration of our EC-QCL set up is shown in Fig. 4.2. The Littman-Metcalf type external cavity configuration consists of four main elements: The gain element (optical amplifier), in this case a quantum cascade laser chip with an antireflection coating on the front facet to inhibit lasing from the Fabry-Perot cavity formed by the front and back semiconductor facets, a collimating lens, a diffraction grating and a scanning galvanometer mirror. The QCL chip is Gold Tin (AuSn) soldered to an aluminium nitride (AlN) carrier, custom-made (Thorlabs Quantum Electronics, USA) with broadband anti-reflection (AR) coated on one facet. The AlN carrier is mounted on a copper heat-spreader with a thermoelectric cooler underneath it. The resulting heat from the thermoelectric cooler is removed by a Corsair Hydro liquid cooler with heat exchanger and standard PC fans to the ambient lab environment. This system is more compact than a traditional chiller and offers an additional advantage of allowing cooling down to -30 °C if necessary. Light is collected with a collimating lens (6.55 mm diameter, EFL = 4.0 mm, with an 8 to 12 µm AR coating on both surfaces). The collimated beam is rotated in polarisation by the two unprotected gold coated mirrors (m1 and m2) and directed onto a diffraction grating

(100 grooves/mm blazed for a wavelength of 10.6 μm). This change in polarisation is carried out because of the highest efficiency of the grating is the orthogonal polarisation to the laser emission. As an example, typical grating efficiency vs wavelength plots are shown in Fig. 4.3. From the grating, the first order diffraction beam is reflected into the QCL chip by the galvanometer mounted mirror. The emitted laser light is extracted in the zeroth order reflection from the diffraction grating. The scanning galvanometer mirror is use for wavelength selection and fast scanning. This cavity configuration is chosen for simplicity of adaptation, and ensures that the output beam does not move in angle, such that a sample scanning and microscopy stage can be easily incorporated. A second benefit is that the double pass on the grating decreases the bandwidth of the external cavity and so increases the instantaneous resolution of the laser. This allows a route to shrink the beam size and grating size.

The laser can run in two modes of operation, pulsed and continuous-wave (CW). In pulsed mode, the gain chip emits light in optical pulses (Appendix C Fig. C.1). The frequency, pulse period and pulse width can be modified accordingly to specific needs. On the other hand, CW operation produces a constant power over a period of time (Appendix C Fig. C.2). Pulsed mode has the advantage of offering larger spectral coverage and requires lower energy compared to CW operation [75, 340]. CW offers narrower line width, it allows the implementation of enhanced modulation techniques which can improve the sensitivity of the measurements, and it has more spectral power compared to pulsed mode. Pulsed operation suffers from pulse-to-pulse intensity fluctuations [341] and in our system it requires a lock-in amplifier, which adds more components and complexity to our prototype.

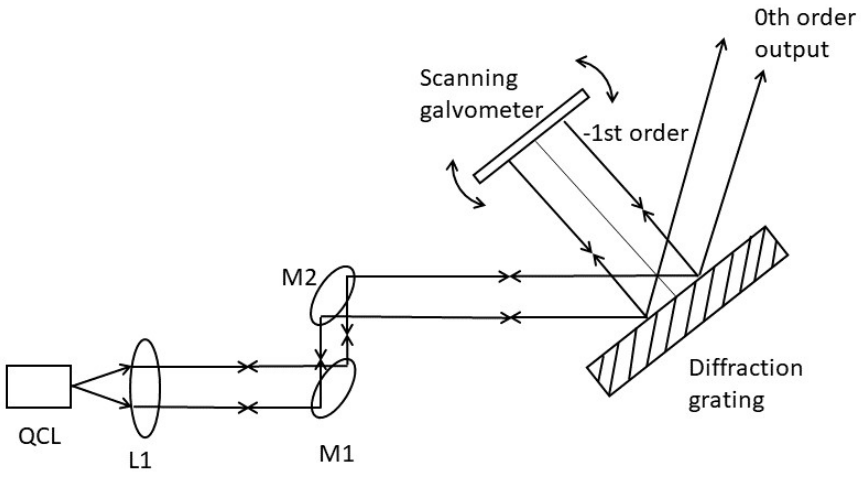


Figure 4.2: Diagram of the external cavity in Littman-Metcalf configuration. The tuning element is a galvanometer scanning mirror, combined with a fixed diffraction grating. Radiation can be extracted from the zeroth-order diffracted light from the grating. M1, 2: gold-coated beam folding and steering mirrors. L1: collimating lens. QCL: quantum cascade amplifier chip

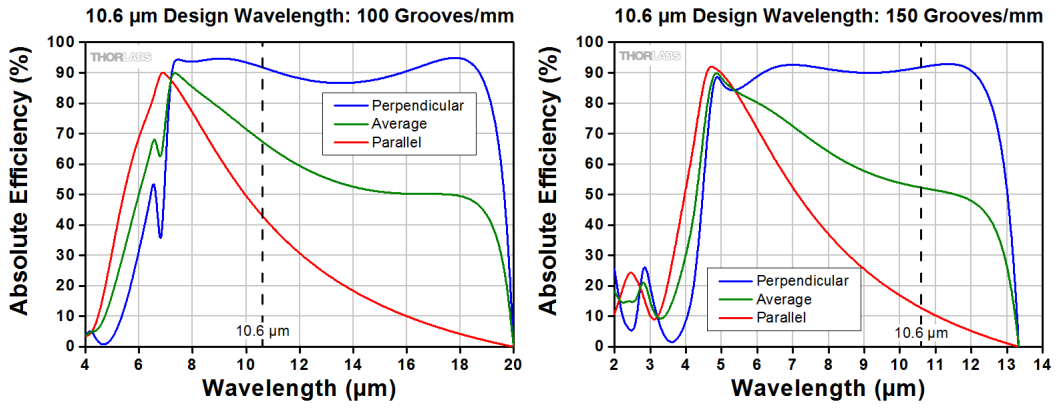


Figure 4.3: Example of grating response to polarised light for 100 grooves/mm (a) or 150 grooves/mm (b).

4.2.2 Scanning system

The galvanometer, Thorlabs GVS111, 1D large beam (10 mm) diameter galvo system with a servo driver board is used to sweep the laser wavelength. The total scanning angle of the mirror is ±12° and repetition frequency up to 1 kHz for small angles. The mirror swings by 0.8° per volt applied from a digital to analogue converter card. The encoder feedback channel from the mirror position is captured by the analogue to digital converter on the same card, along with the detector signal.

4.2.3 Detector

The detector used was an amplified HgCdTe (MCT) photovoltaic detector (2.0 - 10.6 µm, Thorlabs, UK). It has an integrated hyper-hemispherical GaAs lens for optical immersion and a wedged ZnSe window AR-coated for 2 - 13 µm. It has an integrated high gain amplifier which is switchable from 0 to 30 dB. The detector responsivity is shown in Fig. 4.4 and detailed information of detector specifications is shown in table 4.1

Table 4.1: HgCdTe photovoltaic detector specifications

Wavelength range	Responsivity x Effective Detector Width	Output Bandwidth 3dB	Conversion gain p	NEP
2.0 - 10.6 µm (6.5 µm)	≥ 0.01 A·mm/W	DC - 100 MHz	120 V/W (Gain Setting 1) 4000 V/W (Gain Setting 8)	210 pW/Hz1/

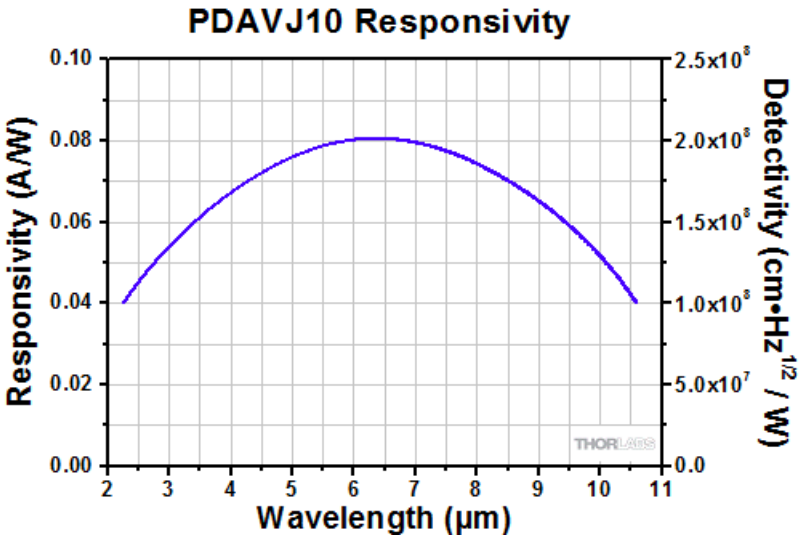


Figure 4.4: Responsivity and detectivity of the MCT detector used in the set-up across 2 - 10 µm wavelength range (taken from Thorlabs, UK)

4.2.4 Electronics and interface

The system is controlled by a two custom-made LabVIEW programs. The first one is use for step scanning and alignment. The LabView graphical user interface (GUI) displays the current position of the mirror and the detector signal simultaneously. The user can set the position of the mirror, maximum and minimum position for scanning (in volts) and the time for each point measure in milliseconds. The result screen displays the detector signal against the mirror position in degrees, and a two-column tab delimited ASCII text file is generate after the scan. The second program is optimised for fast sweep scanning. The output is a waveform that drivers the galvo mirror via the DAQ card. The user can set the number of scans and the speed of the scans. The results display the final spectra in degrees against transmittance. A file with all the scans recorded is generated for post analysis.

All analogue signals from the system are connected to the computer via a DAQ PCIe card (PCIE 1802-AE, Advantech Ltd., UK). This card is a combined multichannel digital to analogue converter (DAC) and analogue to digital converter (ADC). The inputs are the detector signal and the galvanometer position. Analogue signals out of the system are a waveform to the servo driver of the Galvo system. The power supply for the laser and thermoelectric cooler is the ITC-4001QCL bench-top laser diode and temperature controller (Thorlabs, UK). The power source for galvanometer is the GPS011-EC galvo system linear power supply (Thorlabs, UK). The system schematic is shown below (Fig. 4.5). A lock-in amplifier was added when the laser was in pulsed mode. The output of the detector was connected to the lock-in amplifier and, from there, the output was connected to the DAQ PCIe card. A reference signal out of the TEC controller was sent to the lock-in amplifier (Appendix C, Fig. C.3). The alignment of the system was performed using a broadband pyroelectric detector (Gentec-EO, THZ51-BNC), via a beam splitter or more simply with a mid-infrared card (Thorlabs VRC6S).

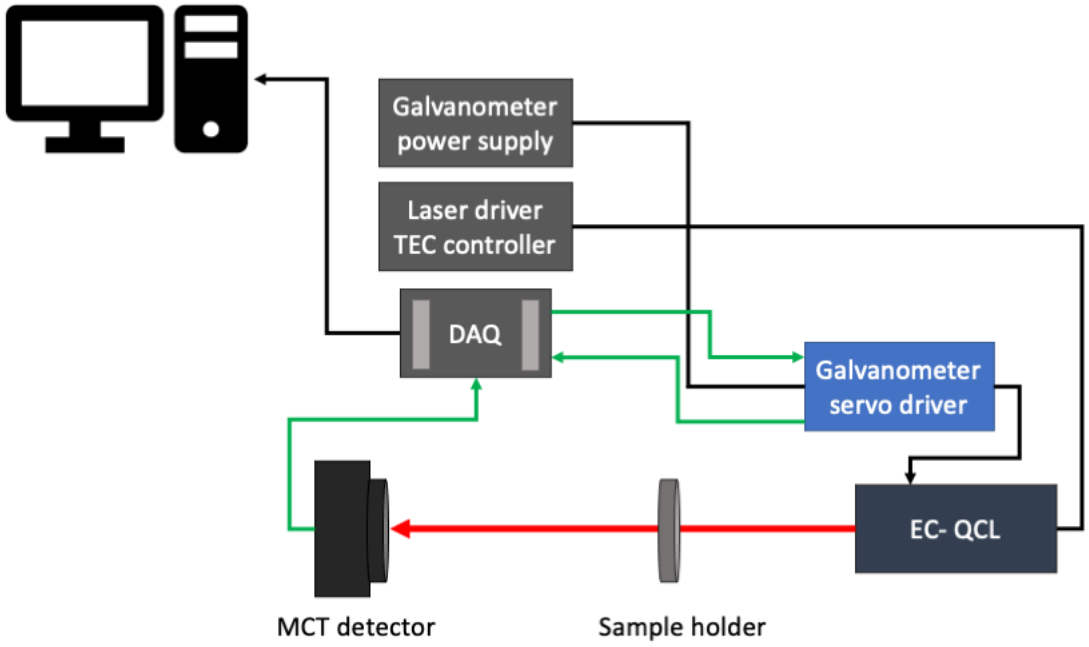


Figure 4.5: Schematic of the QCL-based setup mid-IR transmission for solids

4.2.5 Samples

Various types of samples were used to test the prototype: a calibrated and NIST traceable polystyrene sample (PS) with a thickness of 0.3 and 0.7 µm with characteristic absorption bands located at 1028 cm^{-1} and 1068 cm^{-1}. These calibrated samples were used to compare the spectra collected from the system and a commercial FTIR. Also, they were used to characterise the speed of the system as well as any shifts in wavelength. After that, a set of samples from consumer plastics were used to test the prototype with real world samples and compare to the FTIR. Moreover, the samples were

also used to assess the ability to identify different types of plastic using the spectra collected with the prototype. Finally, to assess the collection of spectra from biological samples, the external layer of an onion and an *Anopheles gambiae* mosquito were used to test the system.

4.3 Results

4.3.1 EC-QCL spectral characterisation

Figure 4.6 shows the emission spectra of the EC-QCL system at different galvanometer angles with a current injection of 1.7 A in continuous wave (CW) operation collected with an FTIR (vertex 70v, Bruker) with a deuterated triglycine sulfate (DTGS) detector, 32 scans with a resolution of 0.3 cm^{-1}. The wavelength of the laser can be tuned over approximately 140 cm^{-1}, from 960 cm^{-1} to 1100 cm^{-1}. The maximum optical output power was 12 mW at 1049 cm^{-1} and output power follows the spectral profile (Fig. 4.8). Instantaneous linewidth measured by the FTIR is 0.3 cm^{-1}, and it was constant across the tuning range of the laser (Fig. 4.8). Spectral purity was characterised by moving the galvanometer at steps of 0.001 volts and collecting the emission spectra at each position. The wavelength varied discontinuously with mode hops. However, the mode hops were not constant, varying from 0.3 cm^{-1} to 0.9 cm^{-1} (Fig. 4.9).

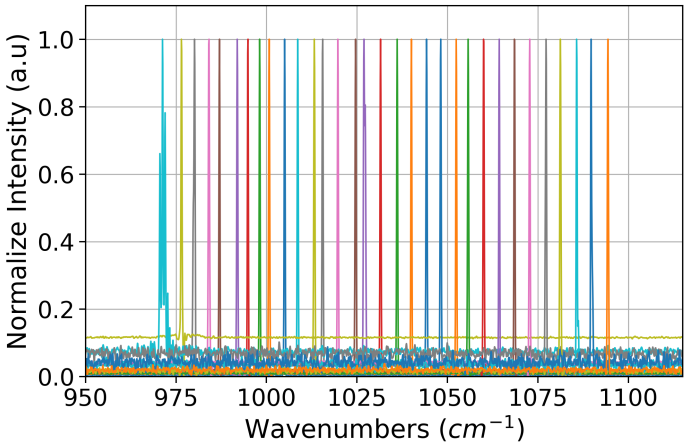


Figure 4.6: Normalised emission lines of the EC-QCL during the wavenumber scan. Each measurement was collected by rotating the scanning galvanometer mirror by 0.1 volts. Current injection of 1.7 A.

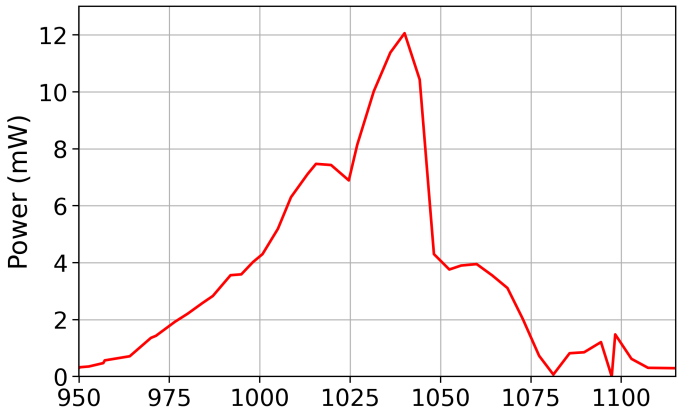


Figure 4.7: Measured CW output power of the EC-QCL as a function of wavenumber with a power meter. The maximum power peak of 12.6 mW was achieved at 1040 cm^{-1}.

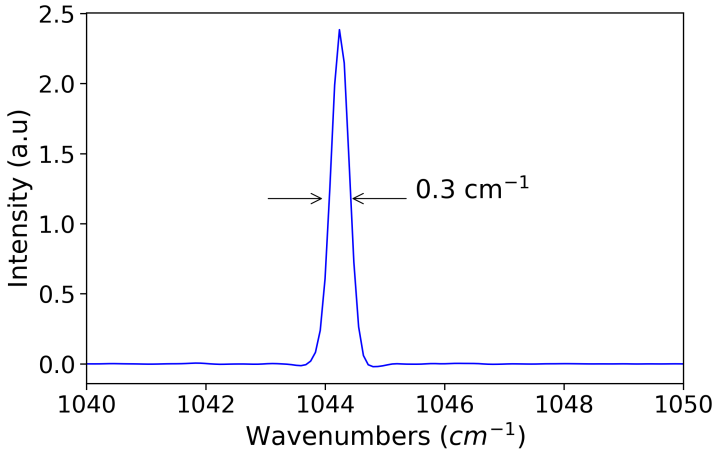


Figure 4.8: Laser spectrum of the EC-QCL at the wavelength of 1044.3 cm^{-1} with a linewidth of 0.3 cm^{-1}.

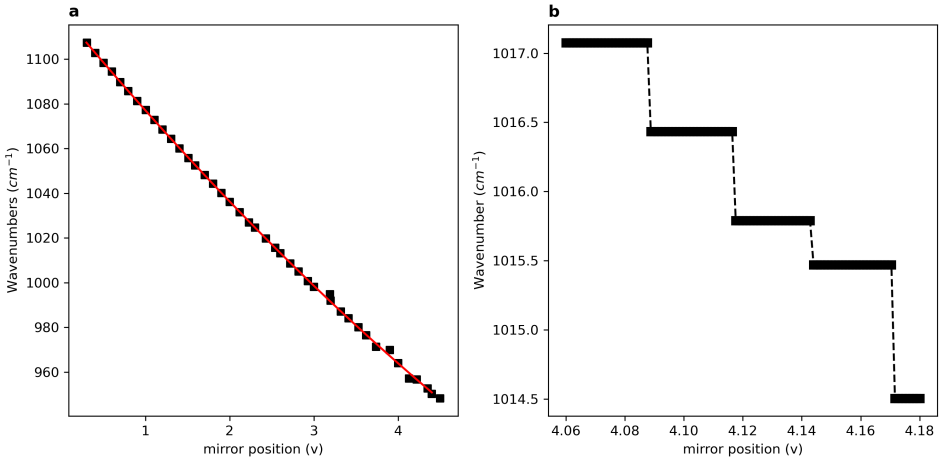


Figure 4.9: a) Step-scan measurements of emission spectrum of the EC-QCL during a scan between 950 cm^{-1} and 1100 cm^{-1} recorded with 0.2 cm^{-1} spectral resolution b) Wavelength measurements at step scans of 0.001 volts. Mode hops range from 0.3 to 0.9 cm^{-1}

Further characterisation was performed by time-resolved spectral measurements of the laser output at 5 Hz with a sine wave from a wave generator driving the galvanometer mirror. The EC-QCL lasing range varied from 1110 cm^{-1} to 950 cm^{-1} following the sine wave movement of the mirror. As a result, one down-sweep and one up-sweep of the tuning range is achieved each cycle, as show in Fig. 4.10.

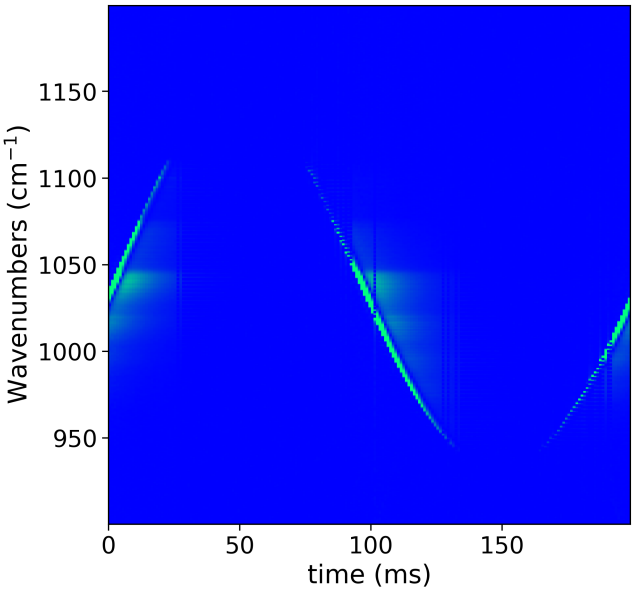


Figure 4.10: The time-resolved emission spectra of the EC-QCL in CW mode with a current injection of 1.7A and temperature of 19 C. Scanning galvanometer is modulated with a sine-wave at 5 Hz

This tuning methodology allows for two spectra per cycle to be captured. At this speed, both sweeps are found to be identical in power vs wavenumber characteristic.

Two different diffraction gratings were tested in the system based on the number of grooves per millimetre. The change from a 150 groove/mm grating to a 100 groove/mm increased the wavelength range by a factor of two from 15.07 cm^{-1} to 30.78 cm^{-1} when moving the mirror over an angle of 1°. The general diffraction grating equation [342] is given by:

$$q\lambda = d(sin\alpha + sin\beta) \tag{4.1}$$

where q is the order, λ is the wavelength of the incident light, d is the groove spacing and α and β are the incident and diffracted angles. In our set-up, q=1, λ = 9.6 µm, d = 10 µm and α=50°, the angle of diffraction β is 11.1°. The diffraction angle is higher when using 150 grooves, which is 42.34°. The change in diffraction grating have a direct effect in the angular dispersion given by the equation 4.2 for Littman-Metcalf configuration:

$$D = 2\frac{\partial\beta}{\partial\lambda} = \frac{4}{\lambda}tan\beta \tag{4.2}$$

By reducing β using a diffraction grating with fewer grooves per mm, the angular dispersion reduces from 21.86 to 4.7° which matches our experimental measurements of ≈ 4.5° (Fig. 4.9a).

4.3.2 Processing EC-QCL data

The fast swept scan mode of the laser required preprocessing due to the absence of a lock-in-amplifier, the high number of scans collected and to eliminate any instrument noise and mechanical errors. One hundred scans were collected for background and polystyrene sample, each scan consisted of 1000 data points. The galvanometer positional values were scanned for duplicates. Approximately, 1% of the 1000 data points were duplicates. Three processing pipelines were tested, two of them were modifications from Schwaighofer et al. [343]. The first one was the co-addition of the 100 scans, the calculation of the final transmittance and the application of a Gaussian filter to reduce overall noise. A Gaussian filter is a low pass filter in which the degree of smoothing can be set by the value of standard deviation of the distribution (notated here as sigma). The second and third pipeline consisted of smoothing each scan using Savitzky-Golay and a Gaussian filter, respectively. To eliminate misaligned scans, the similarity index was calculated for each scan. The similarity index for each scan is calculated by the following equation:

$$similarity\ index = \prod_{i=1}^{100} r|(x_t, x_i)| \tag{4.3}$$

where $r(x_t, x_i)$ is the correlation coefficient between one scan and each of the remaining 99 scans. The values range from 0 to 1 [344]. Scans with similarity index values less than 0.90 for Savitzky-Golay and 0.98 for Gaussian filter were discarded. Approximately, 30 scans were discarded when using Savitzky-Golay and 3 scans when using Gaussian filter. The mean similarity index of the raw data was low (SI = 0.32) but after smoothing, it increased up to 0.99 depending on the filter used (Appendix C, Fig. C.4). After that, the scans were averaged and the final transmittance was calculated by the equation:

$$T = \frac{I}{I_o} \tag{4.4}$$

The use of the first method preserved the range of the laser, and the filter managed to eliminate most of the spectral noise (Fig. 4.11), however, the Gaussian filter reduced slightly the intensity of the peak at 1027 cm^{-1}, when sigma values were high (around 10). In the second pipeline, Savitzky-Golay filter did not reduce the intensity of the peaks, but reduced the range of the measurement. This is mainly due to the use of a high window value (Appendix C, Fig. C.5). Moreover, the noise in the 1050 cm^{-1} was still present. The third method showed a higher noise reduction, high and consistent similarity index values across scans and by keeping sigma value at 5, band intensities were not affected.

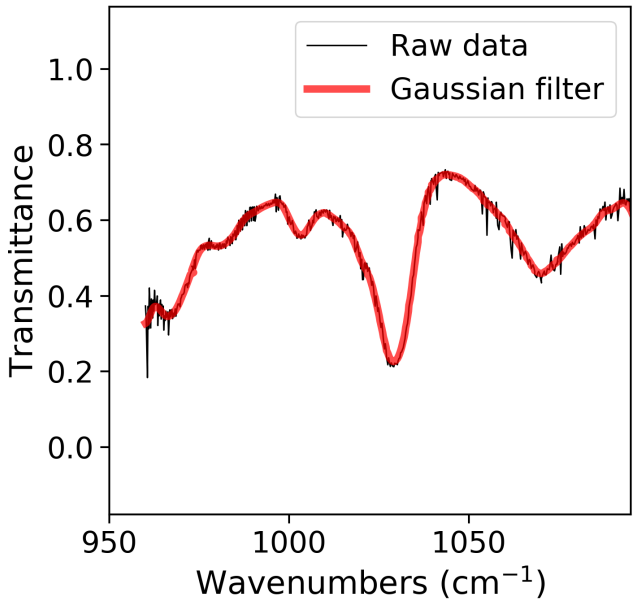


Figure 4.11: PS spectrum collected from 100 co-added scans using EC-QCL (black line) and processed with Gaussian filter (red line).

4.3.3 Speed characterisation

To assess the spectroscopy capabilities of our system at different scanning speeds, the absorption spectra were measured from a calibrated sample made of polystyrene (PS) of 0.3 µm thickness with characteristic absorption bands located at 1028 cm^{-1} and 1068 cm^{-1}. A triangle wave was used to drive the mirror in a linear scan rather than a sinusoidal wavelength sweep as before. This triangle wave was produced by the data acquisition card to the galvanometer at different frequencies ranging from 127 to 500 Hz. Spectral measurements were co-added for an average of 100 scans for each, background and sample measurements, with a total acquisition time of 2 seconds per sample at 127 Hz reducing to 0.8 seconds at 500 Hz. Each measurement consisted of 1000 data points separated by $\approx$ 0.3 cm^{-1}. The two main PS absorption bands at 1028 and 1069 cm^{-1} can be seen at all the different scan speeds (Fig. 4.12). Spectra show high correlation between scan rates, however, there is a slight shift between at faster scanning speeds. The largest shift occurs at 500 Hz of $\approx$ 1.2 cm^{-1} when compared to the 127 Hz. Laser range was also affected by frequency scans, as expected, since the moving mass of the mirror is unable to reach larger deflection angles with a given input power to the galvo motor. It was reduced from 135 cm^{-1} to 80 cm^{-1} when reach 500 Hz (Fig. 4.13). The shift is likely to be caused by a delay due to the rise and fall time of the pyro-electric sensor through the data acquisition card and the more instantaneous reading of the mirror encoder position. The spectrum also shows noise in the 1050 to 1100 cm^{-1} region. This is caused by the lower optical power of the laser, which decreases its S/N ratio.

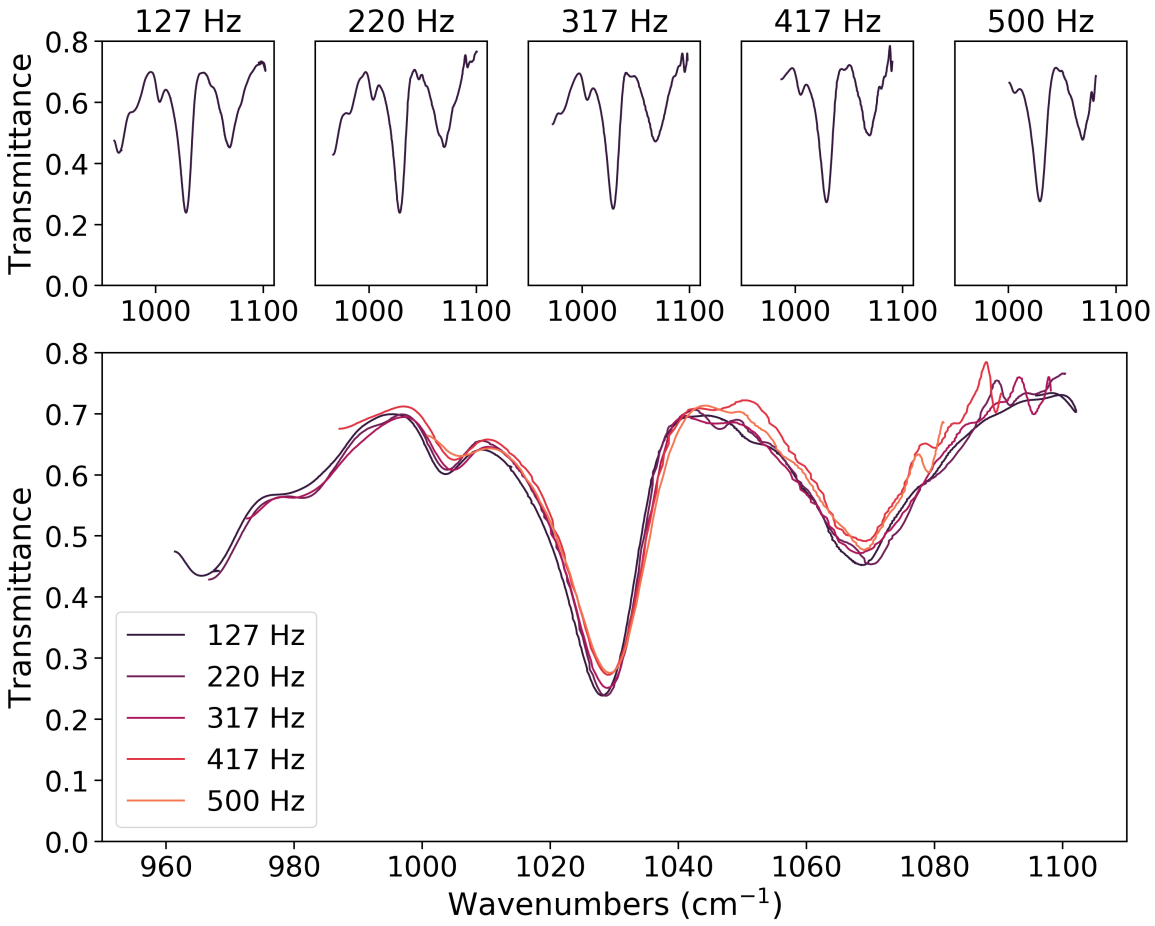


Figure 4.12: Top: Transmission spectra from polystyrene calibrated sample at different scanning frequencies. Bottom: Overlap spectra of PS from 127 Hz to 500 Hz

Figure 4.13a shows the range in cm^{-1} as a function of scanning frequency. As expected, the range of our system decreases as frequency scan increases from 130 cm^{-1} at 125 Hz to 80 cm^{-1} at 500 Hz. While increasing the speed from 127 to 220 Hz or 317 Hz does not reduce the range dramatically (134.48, 133.67 and 125.51 cm^{-1} respectively), reaching frequencies of 417 and 500 reduced the range to 22.76 cm^{-1}. The average optical output power was also measured at different frequency speeds (Fig. 4.13b). By reducing the range of the galvanometer, the optical output power increases from 2.07 mW at 125 Hz to 3.3 mW. This is due to the decrease in angle of the galvanometer at high frequencies, therefore, the mirror spends more time in the centre of the laser range where the optical output power is at its maximum. The highest tuning rate of the EC-QCL in this configuration was achieved with the scanning galvanometer sweeping at 417 Hz, with a resulting tuning rate of 400 µm/s (Fig. 4.13c). Higher tuning speeds would require a lower mass mirror, a higher drive current unit, or a resonant scanning galvanometer, where the tuning speed is fixed to the oscillation frequency of a sprung mirror.

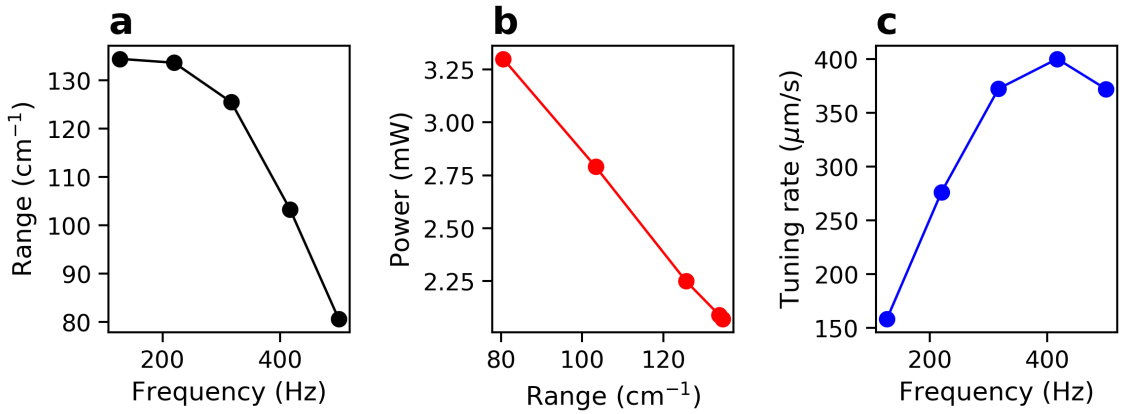


Figure 4.13: a) Tuning range decreases as a function of frequency. b) Average output power decreases as range increases. c) Resulting tuning rate in μm/s at different scanning speed.

4.3.4 Mid-IR Spectroscopy measurements

Polymer identification

Mid-IR spectroscopy measurements were tested first on plastic samples. To assess the agreement between spectra collected from a commercial FTIR and the EC-QCL, a calibrated PS sample and a bottle label sample were used. Spectra were collected using transmission in step-scan mode (Pulse mode, 1.45 A, duty cycle 50%, frequency 5 Hz, pulse period 200 ms, pulse width 100 ms) with a time acquisition of 100 seconds per scan. The spectral features of the two samples can be detected, and there is a good agreement between EC-QCL and FTIR (Fig. 4.14a and b). Polypropylene characteristic main absorption bands at 972 cm^{-1} (CH_3 rock, C-C stretch) and 997 cm^{-1} (CH_3 rock, CH_3 band, and CH bend) [345] were detected with additional bands at 1045 cm^{-1} and 1083 cm^{-1} (Fig. 4.14a). The aromatic CH bend at 1027 cm^{-1} in the calibrated PS sample was also detected and in agreement with the data from the FTIR (Fig. 4.14b). I also compared industrial grading samples of PP and PE against consumer products of the same material. PP and PE samples showed similar absorption bands with differences in transmittance (Fig. 4.14c and d). These differences may be due to local differences in concentration and density across the samples. It should be noted that PE does not have important absorption bands in this region, causing the variability in the number of bands between samples.

After the initial assessment, seventeen plastics samples from consumer products were used with the aim to identify them based on their infrared spectra. Most of the plastic samples were translucent with a wide range of colours. The samples came from labels, covers or containers for food packing and domestic cleaning products. Initial PCA results showed a clear clustering pattern similar to the material indicated by the sample's manufacturer (Fig. 4.16). The first principal component accounts of the 90.5% of the total variation and separate PET from PE. The second principal com-

ponent accounts for 7.9% of variation and appear to separate PE from PET and PP. Polyethylene terephthalate spectra were very similar due to the low transmittance in the 1100 region caused by two absorption bands at 1096 cm^{-1} and 1050 cm^{-1} which are assigned to methylene group and vibrations of the ester C-O bond [346]. Polyethylene samples showed absorption bands at 1050 cm^{-1} and 1080 cm^{-1}, but they were not consistent across samples. Polypropylene samples were also uniform across samples, except a milk bottle cap with high absorption bands at 974 and 999 cm^{-1} (Fig. 4.15). The spectra from the three types of polymer were sufficiently different to from defined clusters.

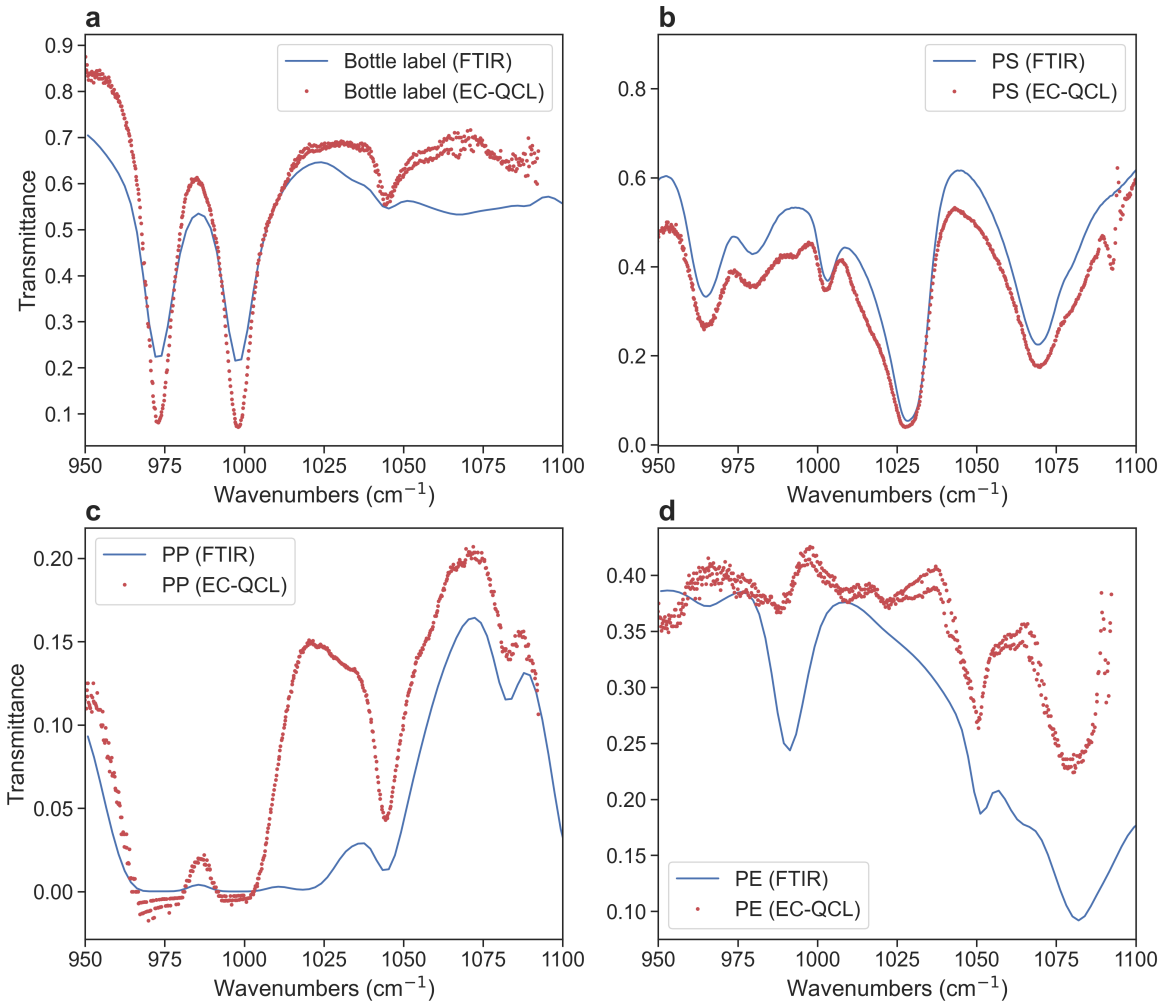


Figure 4.14: Transmission spectra of plastics from consumer products from **a)** bottle label **b)** calibrated PS sample **c)** PP **d)** PE obtained with the EC-QCL (one scan, 200 seconds acquisition time) and with a commercial FTIR (16 scans). Congruence of spectral features between the two systems are comparable.

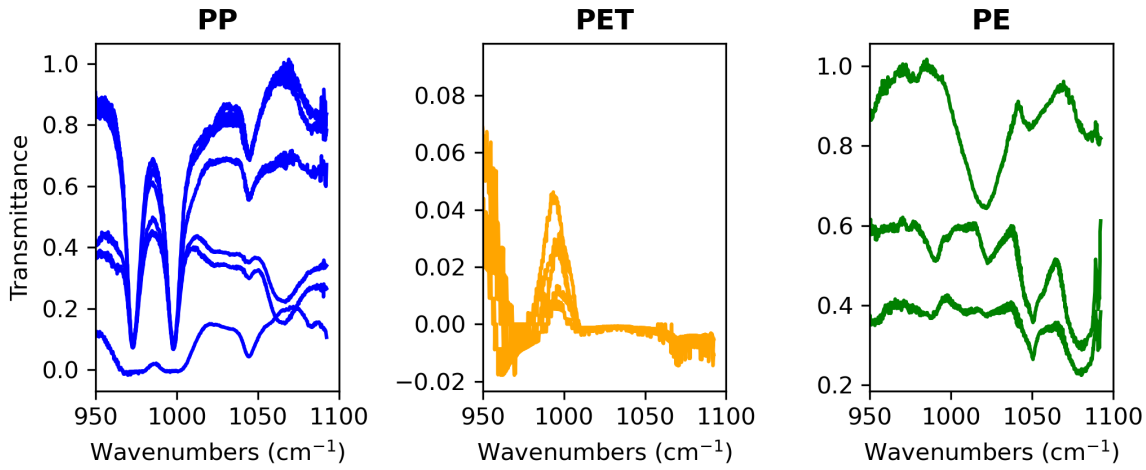


Figure 4.15: Mid-IR spectra of consumer products collected by EC-QCL. Polypropylene - PP (blue), Polyethylene terephthalate - PET (orange) and Polyethylene - PE (green))

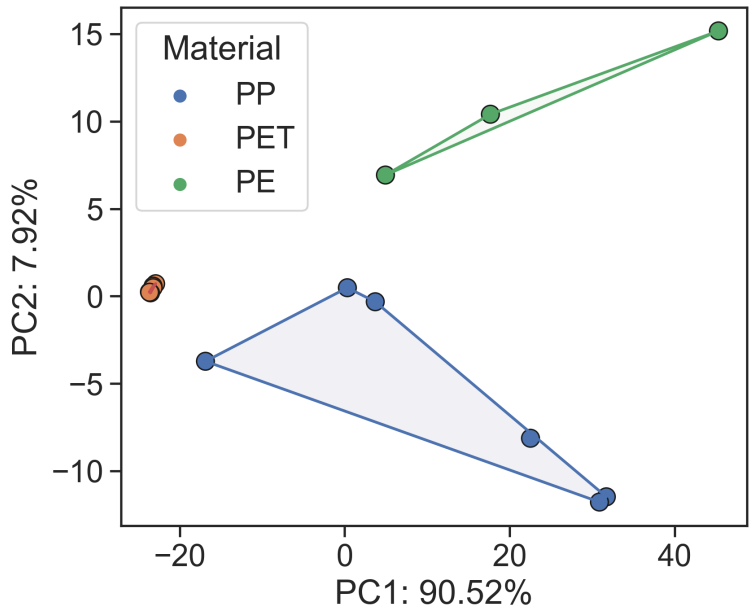


Figure 4.16: PCA scores scatter plots of the 17 consumer plastic products coloured by material: PP (blue), PET (orange) and PE (green). The material of each sample was identified by the manufacturer information.

Fast-swept scans were also tested on the same plastic consumer products. Agreement between FTIR and EC-QCL on the main absorption bands was evident (Fig. 4.17). The EC-QCL spectrum contains more noise since it was unfiltered. A difference in the amount of transmittance in the main peaks at 974 and 999 cm^{-1} and a shift in the baseline is evident.

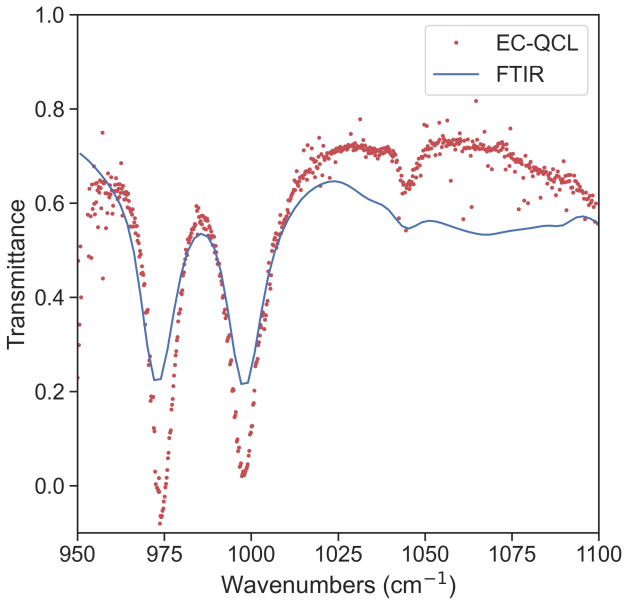


Figure 4.17: Transmission spectra of polypropylene sample obtained with the EC-QCL (acquisition time ≈ 1 second, 100 averaged spectra), and with an commercial FTIR (acquisition time 20 seconds, 20 averaged spectra)

Biological samples

Mid-IR transmission spectra from biological samples were recorded using fast-swept mode. A spectrum was acquired by continuous tuning of the galvanometer across the full spectral range (950 - 1100 cm^{-1}) with a tuning speed of 150 µm/s and an instantaneous spectral resolution of 0.3 cm^{-1} (laser linewidth was 0.3 cm^{-1} at full width half maximum, FWHM). One hundred scans were averaged per spectrum for background and a further for sample transmission for a total acquisition time of ≈ 2 s. The samples used were an external layer from an onion and a whole mosquito (*An. gambiae*). The onion did not require processing (it only needed to be dry) and it was just positioned in front of the laser. The mosquito samples were processed with the KBr disc method for transmission measurements. KBr method consists of grinding the sample with KBr powder. The resulting mixture is then compacted using a press until a translucent disc is obtained. This facilitates transmission measurements. Savitzky-Golay smoothing was applied to the final spectrum and the second derivative was calculated.

KBr is infrared inactive, therefore is widely used as a substrate for transmission measurements since there will not be any bands associated with the KBr disc [347,348]. Moreover, as mentioned before, a background of a KBr disc without the sample was taken to correct for humidity and presence of CO_2. From the onion sample, four absorptions bands were identified at 990, 1014 and 1035, 1056 cm^{-1}. The region of 1800 to 900 cm^{-1} contains bands from cellulose and pectin [349]. The shape of the raw onion spectrum collected with the EC-QCL shows variations in the region 1050 to 1100 cm^{-1} (Fig. 4.18a left). Comparison between FTIR and EC- QCL second derivative

spectrum found agreement with three out of the four absorption bands. The band at 1056 is not well resolved in the EC-QCL spectrum (Fig. 4.18a right). Similar results were obtained with the mosquito sample (Fig. 4.18). Three absorption bands were identified at 988, 1029 and 1056 cm^{-1}. Band 1029 and 1056 are related to chitin [17, 292]. Band assignments are listed in Table 4.2. The shape of the raw spectra is comparable with the FTIR, however, the baseline shift in the 1050-1100 cm^{-1} region is observed. In comparison, between the second derivative spectra, there is an agreement between bands, however, the artefacts of the spectrum from the EC-QCL are amplified by the second derivative method.

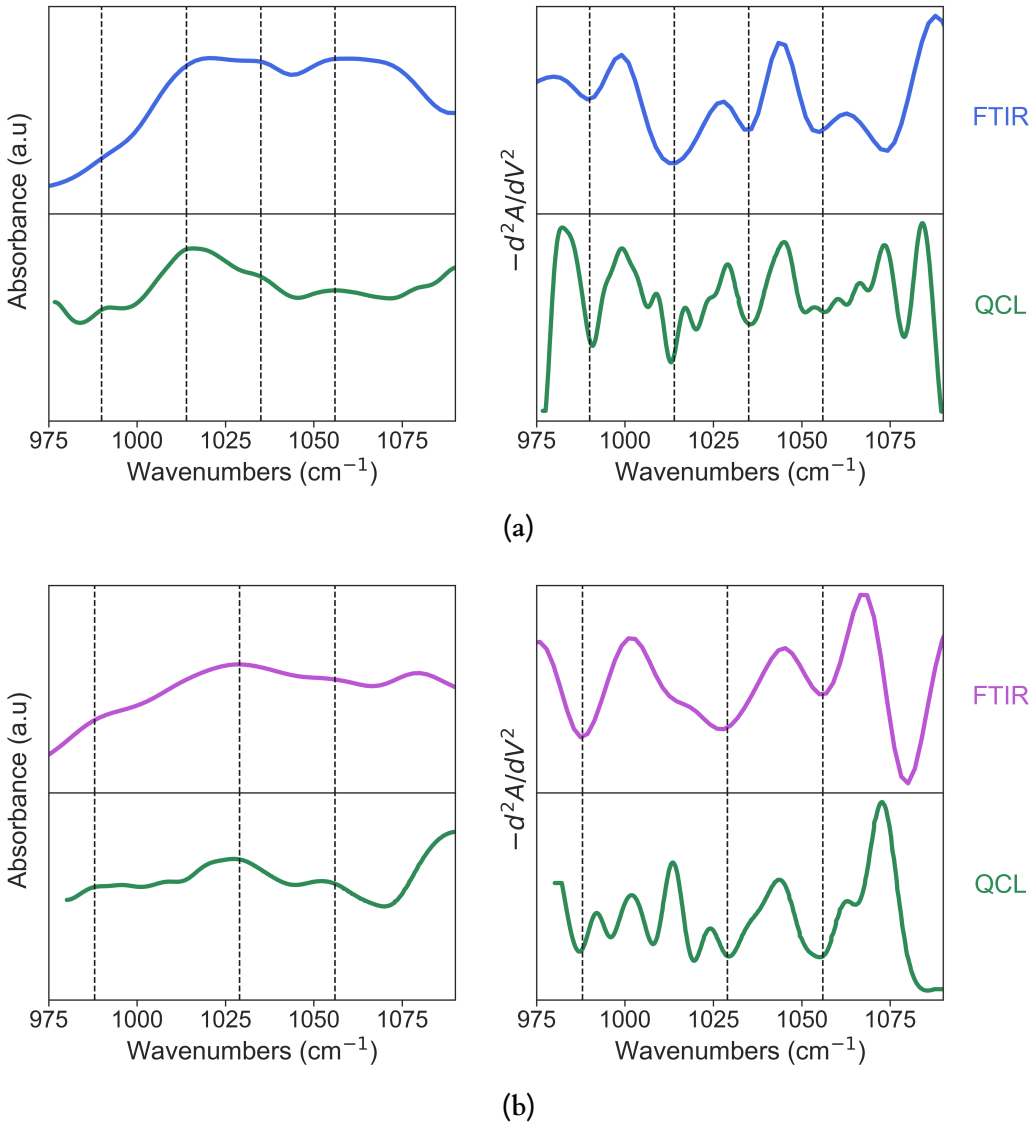


Figure 4.18: Mid infrared spectra of an **a)** onion dried external layer and **b)** *Anopheles gambiae* mosquito recorded with the EC-QCL setup (green line) and with a commercial FTIR (blue line). Black dashed line show the agreement of the spectral features of the second derivative between the two systems.

Table 4.2: Assignment of wavenumber values found and tentative band assignments in mosquito and onion sample using the EC-QCL prototype [17, 292, 349]

Wavenumber (cm^{-1})	Band assignment	Compound
Mosquito		
988	–	–
1029	C-O	Chitin
1056	C-O	Chitin
Onion		
990	–	–
1014	ν(CO)(CO),ν(CC),δ(OCH), ring	Pectinate
1035	ν(CO),ν(CC),ν(CCO)	Cellulose
1056	ν(CO),ν(CC),δ(OCH)	Pectin

4.4 Discussion

Here, I have proposed the use of a fast swept EC-QCL system for mid-IR spectroscopy. The system can scan at repetition rates of between 117 Hz and up to 500 Hz with a spectral range of 135 to 80 cm^{-1} respectively and with maximum tuning rate of 400 µm/s. It has been used to collect spectra from consumer plastic products, thin layers and KBr disc processed samples.

Our system can reach speeds up to 500 Hz with a tuning range of 80 cm^{-1}. As far as we are aware of, our system is the fastest when comparing tuning rate with systems with the same external cavity. Reported EC-QCL speeds with the same external cavity configuration range from 10 Hz for broad tunability (1425-1285 cm^{-1}) [350], 100 Hz (2105-2240 cm^{-1}) [351], 200 Hz (950-1050 cm^{-1}) up to 1 kHz for spectral windows of 1 to 2 cm^{-1} [338]. Systems with different configurations such as Littrow with MOEMS mirror scanners can reach speeds up to 300 kHz and tuning ranges of 300 cm^{-1} [339]. Fast speeds are useful for obvious reasons such as increase the number of samples can be measure but also for real-time gas monitoring, combustion/explosion diagnostics [352] and chemical imaging [334]. At the moment, our EC-QCL can take 100 spectra in approximately 1 second without any optimisation. Even though, the system would require another QCL gain chip to increase the spectral coverage to 1700 to 1500 cm^{-1}, that will increase the scanning time up to 2 seconds per sample. These scanning times can be reduced by decreasing the amount of spectra collected per sample.

The system performed well in step scan mode when collecting polymer spectra. Spectral features

of PP, PE and PET were identified, and spectra were similar to those collected by FTIR. PP specially has strong absorption bands in our EC-QCL region. Moreover, PS is also a good candidate, as shown by the spectrum from the calibration reference sample. PET-labelled samples absorbed most of the infrared light in the 950 - 1100 cm^{-1} region in transmission mode, except for a small region near 999 cm^{-1}. PE has strong absorption bands but does have one small feature at 1050 cm^{-1}. Significantly, the spectra between the three types of plastics are sufficiently different in that spectral region, and that difference can be used for polymer identification. Two studies using QCL have been carried out for this application. Michel, A. et al. [116] used an experimental set-up based on a commercial EC-QCL by Daylight Solutions. They were able to identify PET, high density polyethylene (HDPE), low density polyethylene (LDPE), PP, and PS in diffuse reflection mode in combination with linear discriminate analysis with an accuracy of 97%. I show here that even with a basic PCA analysis, our samples already well clustered by the tree types of plastics used, therefore, I hypothesise that a very high accuracy can be achieved using machine learning algorithms and our EC-QCL system. Moreover, our EC-QCL has a speed advantage of 2 seconds of acquisition time for 100 averaged spectrums compared to 5 seconds acquisition time for 1 spectrum. Apart from fast scanning speeds, the application of plastic identification also needs to consider particle size. For example, samples from seawater contain microplastics with sizes less than 10 µm, which remain challenging for current FTIR systems. However, QCL-based systems have achieved sensitivity to five times more particles under 10 µm [117]. Future work should focus on discriminating other types of plastics such as poly-carbonate (PE), polystyrene (PS), poly(ethylene-vinyl acetate) (EVA) and polyurethane (PU), as well as assessing the particle size sensitivity of the system.

I have demonstrated the use of our custom-built EC-QCL for acquiring spectra from biological samples. The onion was used to assess if there was any shift in the position of absorption bands when comparing FTIR and ECQCL measurements. The chose of an onion was due to its availability, since other biological samples (from other insects) were not available. Both, onion and mosquito measurements were carried out in transmission mode. Due to the thickness of the onion, it did not require pre-processing with KBr. On the other hand, mosquito sample did require sample pre-processing with KBr. Both samples showed a high agreement seen in the position of the absorption bands when comparing with measurements using a commercial FTIR. However, due to the uneven power density of the laser across the sample due to the high spatial resolution of a coherent laser source relative to the incoherent source from the FTIR, aberrations on the spectrum were present. This can be remedied by optimising the power density of the laser to have a more Gaussian shape across a more suitable spots size. Processing the sample into a KBr disc increases the time of sample measurement by at least 1 minute per sample. I chose to process the sample with KBr because the system at that time can only perform transmission mode, which is the easiest to assemble and test. Moreover, the thickness of the mosquito makes it impossible to measure a whole mosquito in transmission mode, which is why it needs to be processed into a

KBr pellet. However, QCLs can be adapted for diffuse reflection and ATR [353–355]. As shown in chapter 2 and by Sroute et al. [224], DRIFT is a viable alternative to collect mid infrared spectra from mosquitoes. The implementation of a DRIFT set up is fairly easy (one is already built in the laboratory but due to COVID-19 lock-downs there was no time to collect enough data to present in this thesis) and future work should focus on testing the performance of the sampling technique with whole mosquitoes and EC-QCL. Examples of high performance EC-QCL using diffuse reflectance can be found in analysis of pharmaceutical formulation [356, 357]. Using this approach allows non-contact and non-destructive measurements. The range of our system is quite small compared to the 400-4000 cm^{-1} range of commercial FTIR use on previous studies in *Anopheles* mosquitoes [17, 229] and for malaria diagnosis [230]. However, not all the mid-infrared region is needed for good discrimination. For example, species identification in *Aedes* has been carried out using only the 1700 to 600 cm^{-1} region [224]. In malaria diagnosis from blood samples, most of the model coefficients were located at 1100 to 800 cm^{-1} region [230]. Even though the limited range of the QCLs can be seen as a disadvantage, not having to record the whole spectrum makes QCL system more efficient and faster compare to FTIR. This benefits not only for large volumes of spectral collection but also for future applications such as chemical imaging. If increment of spectral range is needed, gain chips with multiple-stage hetero-cascading active regions, can increase the spectral range to 772 cm^{-1} [87]. Also, multiple QCL gain chips with different wavelength ranges can be combined, extending the range accordingly to specific needs. Commercial systems have up to 4 gain chips for a coverage of 5.4 - 12.8 µm (800 to 1800 cm^{-1}) [358]. Therefore, there is great potential of EC-QCL for mosquito surveillance thanks to its versatility and customisable nature.

The noise at the 1000 cm^{-1} region of the system is due to the uneven spectral power density. Usually, systems show a Gaussian-type distribution of power, however our gain chip has a very steep increment in power in the mentioned range. Therefore, not all the wavelengths have an even power. This is more evident when running the system at full speed, where the overall power decreases and its signal-to-noise ratio also decreases. This can be overcome by improving the alignment of the system and by miniaturisation onto a more stable micro-bench rather than the large traditional optical breadboard. Slight misalignment between the galvanometer and the diffraction grating can cause loses in power and range. Another way of removing noise is by pre-processing the signal. Averaging scans can help to remove noise, but it cannot eliminate it if the noise is not random [359]. A variety of filters such as Savitzky-Golay, and Gaussian filter utilised here showed and improvement in the signal but other filters such as Fast Fourier Transform (FFT) can help. The addition of similarity index here helped to avoid the use of high values of sigma by removing misaligned scans. This technique has been proven useful to increase S/N ratio and sensitivity [53, 343]. Another way to improve S/N ratio is adding balanced detection. Balanced detection consists of splitting the laser signal by a beam splitter or beam sampler into a reference signal and sample signal. Both signals are detected with matched detectors, and the

reference signal is subtracted from the sample signal in real time. This process effectively cancels all electrical noise [360] at least up to the sample plane. This method can reduce noise by 20 times [53], however, the use of two detectors increases the cost and complexity of the setup and limits portability. Therefore, a compromise between portability, noise and cost needs to be addressed before choosing a given spectroscopy system going forward.

4.4.1 Future work

Future work should focus into three main areas: increasing wavelength range or move to another mid-IR region, diffuse reflectance as sampling technique and increasing scanning speed. For the wavelength range, even though 950-1100 µm is an important region for biological applications, the laser range can be extended by at least 100 cm^{-1} accordingly to the gain chip manufacturer. Moreover, the 1700-1500 cm^{-1} has shown to be important for age grading mosquitoes [222,310]. Therefore, acquiring gain chips in the amide I region should be put into consideration. Second, the main motivation of optical techniques is the non-contact non-destructive characteristic to measure samples. Focusing on development of a diffuse reflectance setup is key. Even though, thorax and abdomen are high absorbance tissues due to the thickness, high power lasers can overcome this issue. Diffuse reflectance also can increase the high-throughput nature of the system, by reducing sample processing time. There would be no need for KBr pellet processing or positioning the sample under the diamond crystal as in ATR. Moreover, automatisation of sample measurements by incorporating carousel-type holders can greatly increase the amount of samples that can be measured in a given time. Finally, the speed of the laser sweep can also be further enhanced by changing the mirror scanning system. Options outside traditional galvo systems are oscillatory scanners (MOEMS and resonant mirrors) and rotary (polygon) scanning systems. QCL systems with MOEMS diffraction gratings can achieve speeds up to 975 Hz with a broadband range in the Littrow configuration [339]. Resonant mirror scanners are also a suitable option for high scanning speeds. Off the shelf resonant scanners can achieve 8 kHz to 12 kHz [361]. However, there is no full control of position or scanning speed compared to traditional galvo scanners. Polygon mirrors, or multi-facet mirrors, offers fast scanning speeds and relative compactness. Through the use of such polygon mirrors, fast-swept systems in the near infrared targeted at optical coherence tomography (OCT) have increased scanning speeds to 50 kHz [362] and 86 kHz [363] using polygon mirrors. The inclusion of these systems can greatly increase the overall speed of future prototypes.

Finally, the prototype at this stage cannot be used in the field. The modifications needed to test it in the field would be: the reduction of the external cavity length, by changing the XYZ stage and eliminating the 3 gold mirrors, so the light will hit directly to the diffraction grating. This will eliminate the difficulty to align the laser and will increase the laser power, increasing its S/N and probably the range of the gain chip. By reducing the size of the external cavity, a small diffuse

reflectance module with a small carousel-type holder can be mounted in the same breadboard along with the detector. In addition to that, an enclosure to protect the system, changing the PC for a laptop and an adding a small USB DAQ can enhance even further the portability of the system. These fairly easy changes would make the system ready to be tested outside controlled conditions.

4.5 Conclusion

In this chapter, I have demonstrated the use of a fast swept continuous wave EC-QCL system for mid-IR spectroscopy in mosquito samples and polymers in the region of 950-1100 cm^{-1}. In summary: i) High speed and high tuning rate were achieved by using off-the-shelf optical components, which also means that there is plenty of room for future optimisation. ii) The system can collect spectral information from polymers in consumer products (PP, PE, PET) and PCA cluster samples according to polymer type. iii) Spectral data from biological samples, onion's external layer of whole *An. gambiae* processed in KBr disc were obtained at high speeds of 100 spectra per second with little optimisation. iv) Absorption bands in the spectrum from mosquitoes matched their position comparing with commercial spectrometers. The current prototype has the potential for extreme miniaturisation, and also lower cost for its use in the field.

Chapter 5

General Discussion

5.1 Overview

Mid infrared spectroscopy has the potential to become an additional tool for mosquito surveillance. Studies have shown promising results when using laboratory, semi-field and field samples for species identification and age grading in *Anopheles* and *Aedes* mosquitoes [17, 222, 224, 310]. Furthermore, advances in mid-IR light sources have opened the development of faster and better spectrometers with increased portability and miniaturisation, easy integration, and lower cost. These characteristics are appealing for the development of a field-ready spectrometer for real-time mosquito surveillance. Therefore, I developed a fast swept QCL-based spectrometer using the 8-11 µm region, aiming for age and species prediction. Due to the ability of QCLs to produce high power mid-IR light and focus it into a small spot size, I investigated the viability of collecting spectra from different mosquito parts using µDRIFT and their use for species, age and cuticular insecticide resistance prediction. Moreover, I identified the most informative region for species prediction from ATR derived data using supervised machine learning to implement future QCLs. Understanding the potential of different tissues and their use with µDRIFT for mosquito surveillance, what wavelength regions are important for efficient data acquisition and how semiconductor lasers can be integrated for next-generation spectrometers, can help to expand MIRS into a routine tool in mosquito monitoring. Here, I review the principal findings and their implications with respect to the use of infrared spectroscopy for mosquito surveillance.

5.2 Principal findings/Implications

This study has made the first step towards the use of QCL-based system for spectroscopy of insect disease vectors. The fast swept EC-QCL system developed here from off-the-shelf components

can already achieve fast scans at speeds from 127 Hz to 500 Hz, with a maximum tuning rate of 400 µm/sec. A sample's spectral profile can be obtained in ≈ 2 seconds, collecting 100 scans of a background and the sample when using at 127 Hz. As the main application of this system was for mosquito spectroscopy, the prototype was able to collect mid-IR spectra from mosquito samples processed into a KBr disc in transmission mode with high agreement in the position of absorption bands compared to a commercial FTIR. However, there are still optimisation steps needed to correct distortions of the spectra and decrease overall noise. In addition, the system needs to be adapted for diffuse reflectance measurements to avoid process the samples. Overall, our QCL system is suitable for fast-swept MIRS of polymers and mosquitoes in KBr pellets. Quantum Cascade Laser-based systems have been tested through a wide range of applications. With the increase in the number of companies building gain chips and the commercial availability of ready-to-use systems for spectroscopy and micro spectroscopy, this technology will become easier to test in vector research. Now that our understanding is expanding in terms of how infrared data from the mosquito cuticle can be related to biological traits, it will become increasingly straightforward to adapt QCL technology to already identified needs, (i.e., custom gain chips to cover discrete important regions of the infrared, automated sampling, smaller and portable spectrometers). This is happening already with bulky and high maintenance systems such as µFTIR that have been replaced with commercial QCL counterparts which are a fraction of the size, with faster data acquisition speeds, and without the need of LN_2 cooled detectors. Quantum cascade laser-based spectrometers are an attractive alternative to expand the area of spectroscopy to disease vectors.

As QCLs have a limited wavelength range, I tested how choosing smaller regions of the infrared window affects species prediction accuracy. When using laboratory samples, a region of 200 cm^{-1} in length located at 1700 cm^{-1} was enough to distinguish between *An. gambiae* and *An. coluzzii* with very high accuracy (> 0.98). The same tendency of different windows having different accuracies was also found in other data sets. Overall, smaller windows of ≈ 300 cm^{-1} located in the Amide I and II region or the combination of two windows can be used to achieve relatively accurate species prediction in *Anopheles*. This goes in line with previous studies showing that not all the spectrum is needed for predictions, and that specific regions are more important than others [222,229]. These results can be used to find more efficient ways to collect data, increase sampling speed and explore QCL based systems with narrower wavelength ranges available compared to FTIR equipment.

This work adds evidence of the use of legs for species and age prediction in mosquitoes, and for the first time in *An. gambiae*. Infrared spectroscopy is on its way to become a complementary tool for mosquito surveillance, due to its straightforward implementation, the minimal training needed to operate the equipment, and its high throughput at a lower maintenance cost compared to molecular techniques. Current FTIR studies in mosquitoes rely on ATR as a sampling technique to extract spectral information [17, 199, 222]. Attenuated Total Reflection is a robust technique for solids, however, one of the disadvantages is its destructive nature to the mosquito

sample. Moreover, the requirement of pressing the sample against the crystal limits the sampling speed and challenges any possibility of automated sampling. Additionally, collecting localised spectra from specific body parts such as legs or wings is not possible due to the small amount of tissue these parts have. Here, I demonstrated how the use of μDRIFT can overcome these limitations. First, by measuring the backscattered light from the sample, the destructive nature of ATR is eliminated, and its sampling speed is increased. Second, by using a microscope to focus the infrared light onto smaller spots, spectral information from legs, wings or any section of the mosquito can be extracted. Age prediction into categories of 3 days or 10 days old was possible (mean accuracy of 77.1% ± 6.5%) and, to a lesser extent, species prediction between *An. gambiae* and *An. coluzzii* (mean accuracy = 69.1% ± 7.9%). Age prediction accuracy is lower compared to NIRS studies [214, 217]. However, by increasing training sample size, model performance is likely to improve. I also investigated whether this method could distinguish between 3 mosquito strains from two species (Kisumu, Tiassale and Ngousso) that vary in cuticular resistance to insecticides. Preliminary results showed strains could be differentiated into those with (1 strain) and without cuticular resistance (2 strains) with an accuracy of 71.3% ± 8%. This raises the possibility that some important insecticide resistance phenotypes could be identifiable using MIRS; However, further work is needed to test this as the preliminary results here could be confounded by other factors such as the use of different strains. The relative high accuracy for age grading opens the possibility of testing the hypothesis that different tissues might be better suited for the prediction of specific biological traits. This can help us improve our understanding of how distinct morphological tissues contain different information for different biological traits. Previously only one study has used μDRIFT to analyse mosquito legs for species prediction, in *Aedes* [224] and two others used this approach for specific analysis of insect wings [292, 364]. The successful demonstration of this approach on *Anopheles* legs highlights this approach could be broadly transferable to a range of different insects and vector species such as Tsetse flies (*Glossina* genus), where morphological methods (ovary cycles) cannot be applied to male. Gene expression assays have shown changes across age of genes related to cuticle proteins in Tsetse [365]; further highlighting the potential value of MIRS for age grading in these species. This will open opportunities for new lines of research to characterise disease vectors (or other insects) using MIRS.

This work also revealed how PLS algorithms are still an option to use for prediction of mosquito species and age using MIRS data. PLS had similar accuracy for species prediction and age grading compared to other machine learning models when using data from the same origin. It also highlighted how PLS and other machine learning algorithms could not generalise enough to predict unseen data from other laboratories and from different rearing conditions. The main cause for this was the mismatch (i.e., spectra between data sets are vastly different between training and test sets) rather than the algorithms used. This was apparent when samples clustered together accordingly to the rearing conditions when using PCA. Also, the model coefficients for species classification emphasised different regions of the spectra depending on which data set was used

for training. This creates the necessity of having training data sets from a variety of places and having them updated regularly to have training data sets that match the real-world field samples and have high prediction accuracy. This practice is usually used in state-of-the-art image recognition, where models are trained with billions of samples [366].

5.3 Limitations/Future work

It is important for future development of the QCL prototype to focus on overcome the limitations of the system. First, using KBr treated samples is not optimal, therefore, the exploration of other methods for spectral acquisition techniques needs to be tested. A diffuse reflectance set up was already built at the Photonics Devices and Systems laboratory and preliminary measurements were performed on whole mosquitoes, however the process was interrupted by COVID-19 lockdowns and the results were not included in this thesis. The prototype diffuse-reflectance module can benefit of the development of an automated sample platform to increase sample processing. Second, the addition of a gain chip that covers the amide I and amide II region should be added due to its importance for species and age prediction. The gain chip used was already bought before starting this project and the range was chosen based on previous studies on glucose monitoring and not on studies of MIRS in mosquitoes (which were not published at that time). The current range is useful not only for glucose monitoring [54, 109], but for ethanol and breath acetone measurements [113], malaria detection in blood [228, 229] and urea [228]. However, this region proved to be challenging to assess mosquito cuticle due to the absence of very strong absorption bands. Third, the use of quantum cascade detectors (QCD) to overcome the issue of saturation. TE-cooled MCT and pyroelectric detectors in the mid infrared region have a low power dynamic range (few mW) which causes saturation effects when using high power QCLs. Most of the QCL-based systems need to attenuate the power output of the laser with sapphire windows or a mesh to overcome the detector saturation, but this hinders the advantage of using high power QCLs. Quantum cascade detectors [367] show a wider dynamic range compared to TE-cooled MCT or pyroelectric detectors, allowing the use of the full power of the QCL gain chip. This type of detector has been already used in gas sensing systems [368, 369] and liquids [370], showing similar performance compared to LN_2-cooled MCT detectors. All these options and limitations need to be considered in future developments of the prototype. Once all these changes are implemented, the next step should be reducing the size of the system. Compactness of the system can be achieved by using custom EC-QCL from companies such as Fraunhofer IAF and Fraunhofer IPMS [90, 339, 371] which uses resonant micro-opto-electro-mechanical system (MOEMS) for the external cavity. The use of MOEMS can reduce the size of our prototype by a fifth of the current size. All these future works will take the prototype closer to the ultimate goal of having a fast field friendly mid-infrared spectrometer for mosquito surveillance.

It is worth exploring how the spectral data of each tissue of the mosquito can be used for age grading, species identification and other biological traits such as insecticide resistance or malaria infections with more detail. Currently, abdomen and thorax samples are mainly used in spectroscopy analyses of mosquitoes. By using μDRIFT, the analysis of more localised, specific tissues (i.e., head, thorax, abdomen, legs, and wings) is possible. I recommend testing μDRIFT with different species (i.e., *An. gambiae, An. arabiensis, An. funestus*) and for age grading, more fine scale resolution of age groups should be included (i.e. 1-2 days) and validate the results with parity status to assess the robustness and accuracy across species and ages. Wing samples may be more tractable for analysis given the relative ease with which spectra can be collected from them; however, more detailed investigation is needed to confirm what biological signatures are detectable from wings. from which might be interesting to see what information they may be able to provide.

Insecticide resistance was difficult to evaluate using MIRS due to the experimental design used, and the confounding factors discussed in chapter 2. The use of the same strains known to have cuticle resistance, but that have lost it over time, could be used to disentangle the strain-specific confounder. Also, it will be useful to combine spectral data with hydrocarbon concentration and cuticle thickness determined by electron microscopy to know if the changes in spectrum are due to changes in cuticle composition or are the inherent differences of each strain used.

I hope the work described in this thesis can guide new ways of thinking about spectroscopy in mosquitoes and other insects vectors, and what semiconductor laser-based systems can offer to improve and move the technique forward.

Appendix A

Chapter 2 Appendix

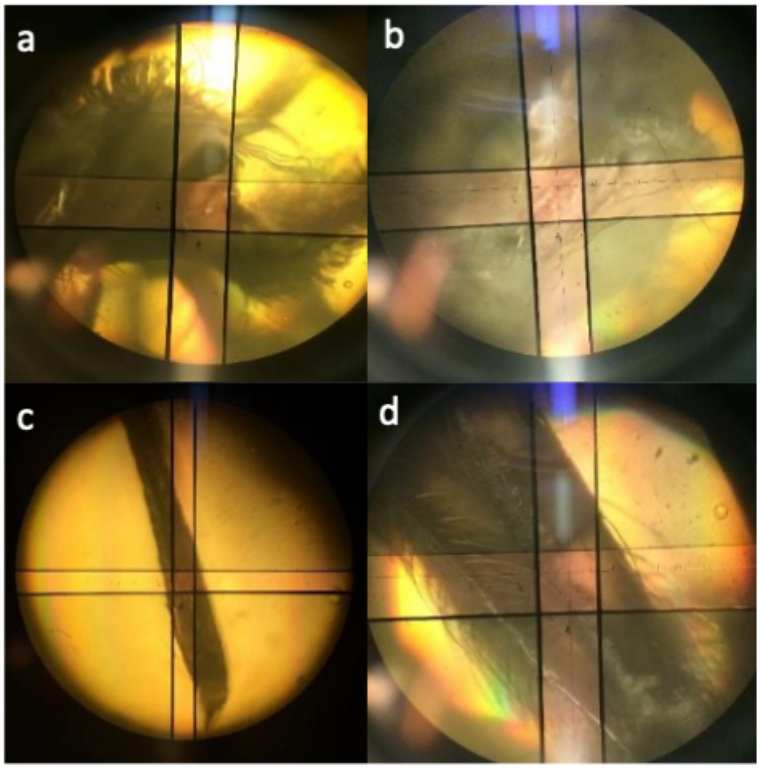


Figure A.1: Preliminary Field of view from μDRIFT in visual mode for a) Head, b) Thorax, c) Leg d) Abdomen

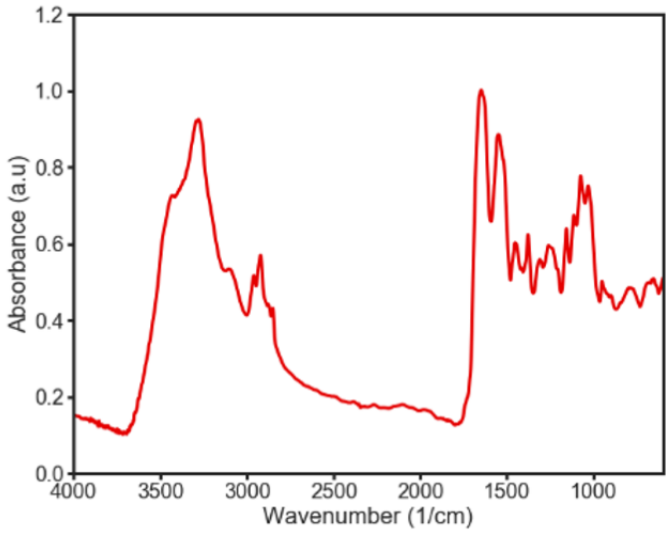


Figure A.2: Leg spectra using μDRIFT with a range of 4000 cm^{-1} to 600 cm^{-1}

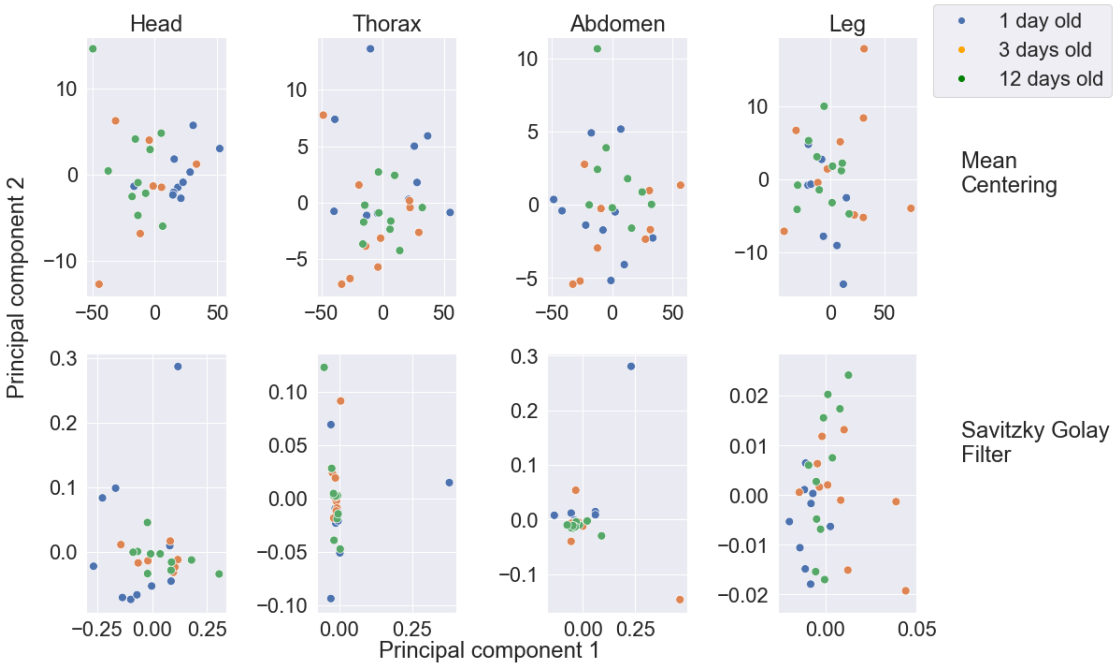


Figure A.3: PCA scatter plot of spectra for each mosquito part on different ages, one, 3 and ten days old. Two different pre-processing algorithms were tested to increase clustering between groups

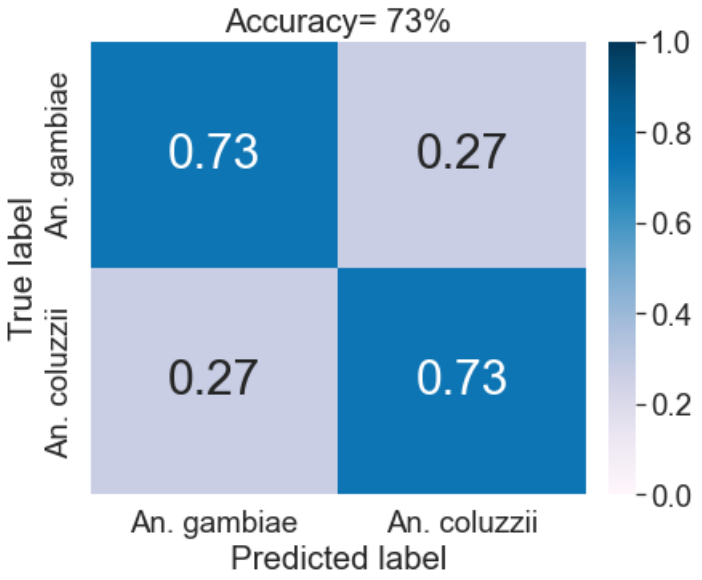


Figure A.4: Confusion matrix of random forest classifier evaluated on hold out set for species prediction.

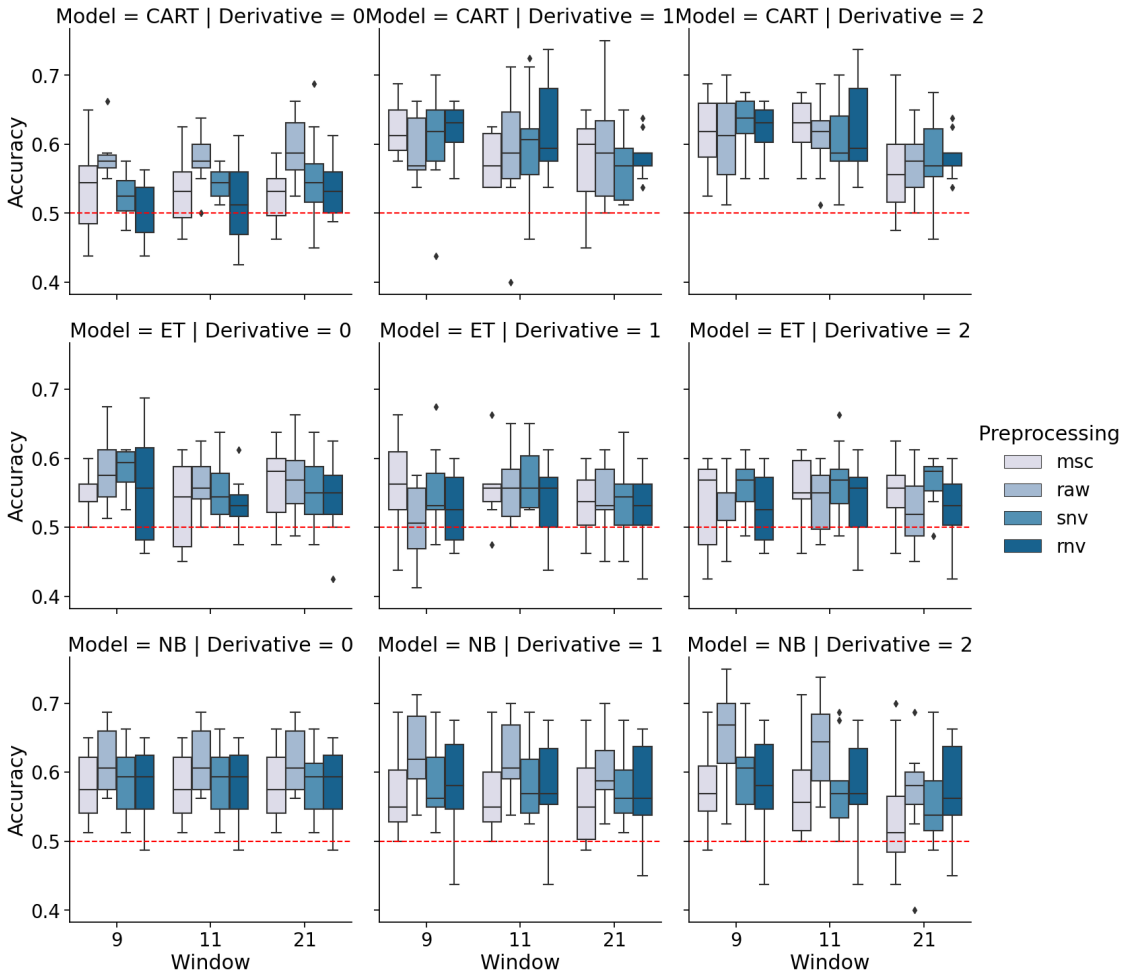


Figure A.5: Accuracy of all models with the different pre-processing algorithms for species prediction

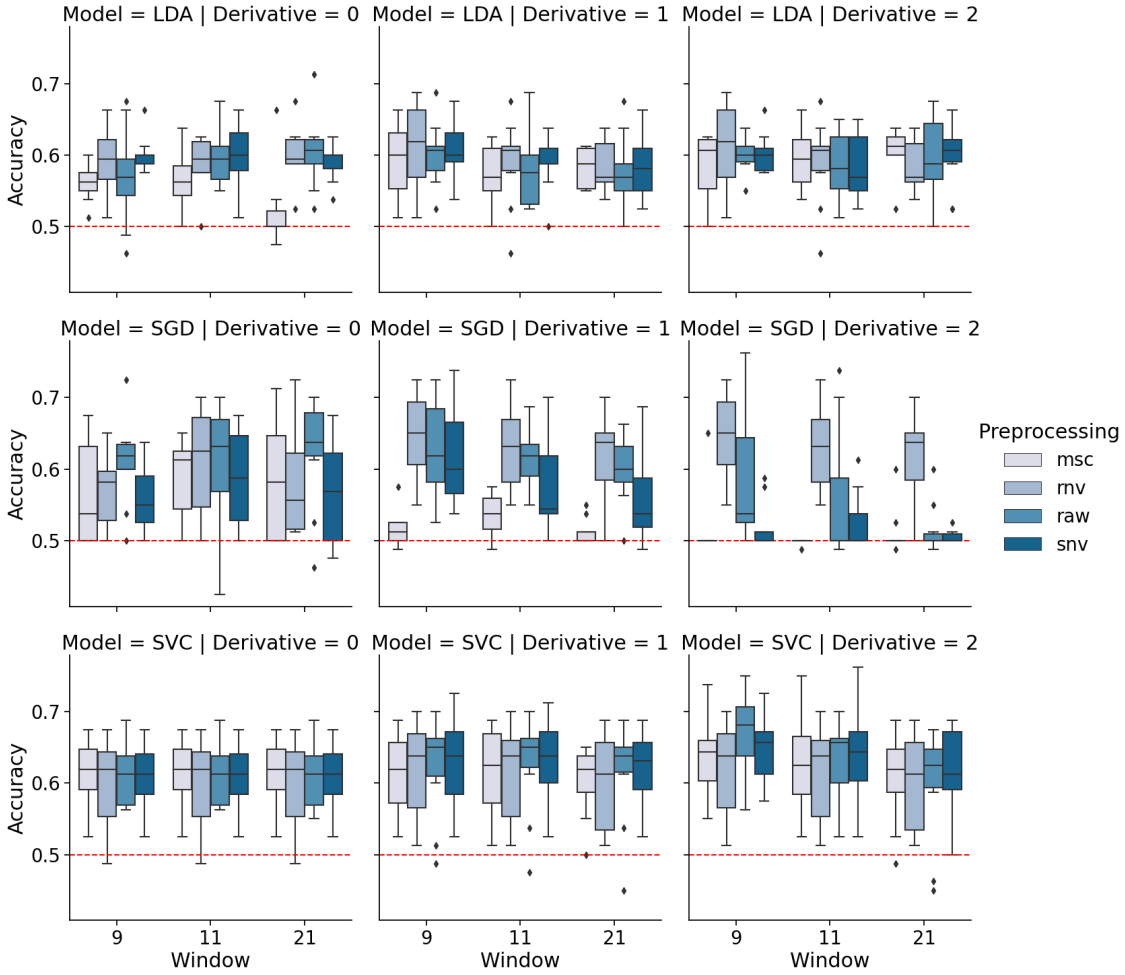


Figure A.6: Accuracy of all models with the different pre-processing algorithms for species prediction

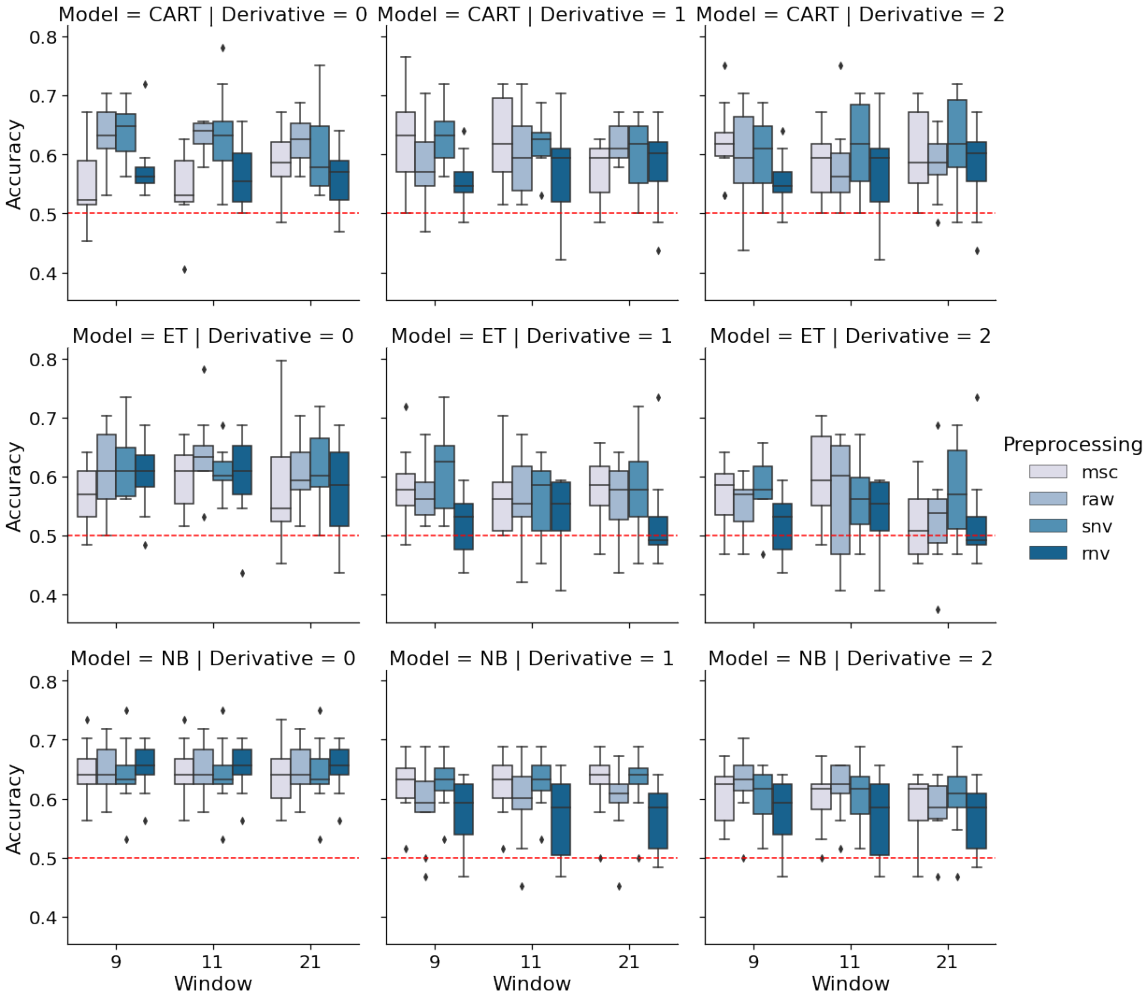


Figure A.7: Accuracy of all models with the different pre-processing algorithms for age prediction

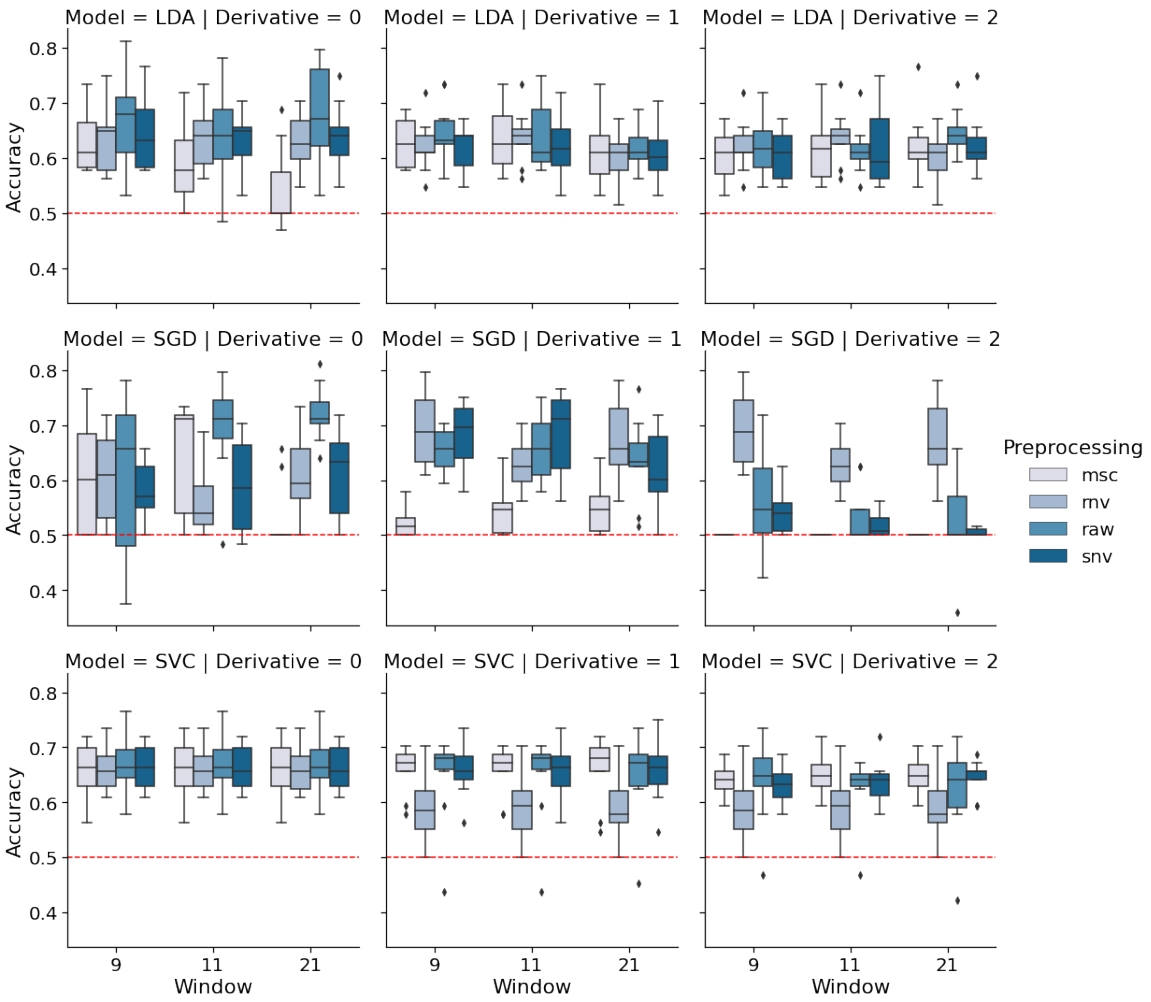


Figure A.8: Accuracy of all models with the different pre-processing algorithms for age prediction

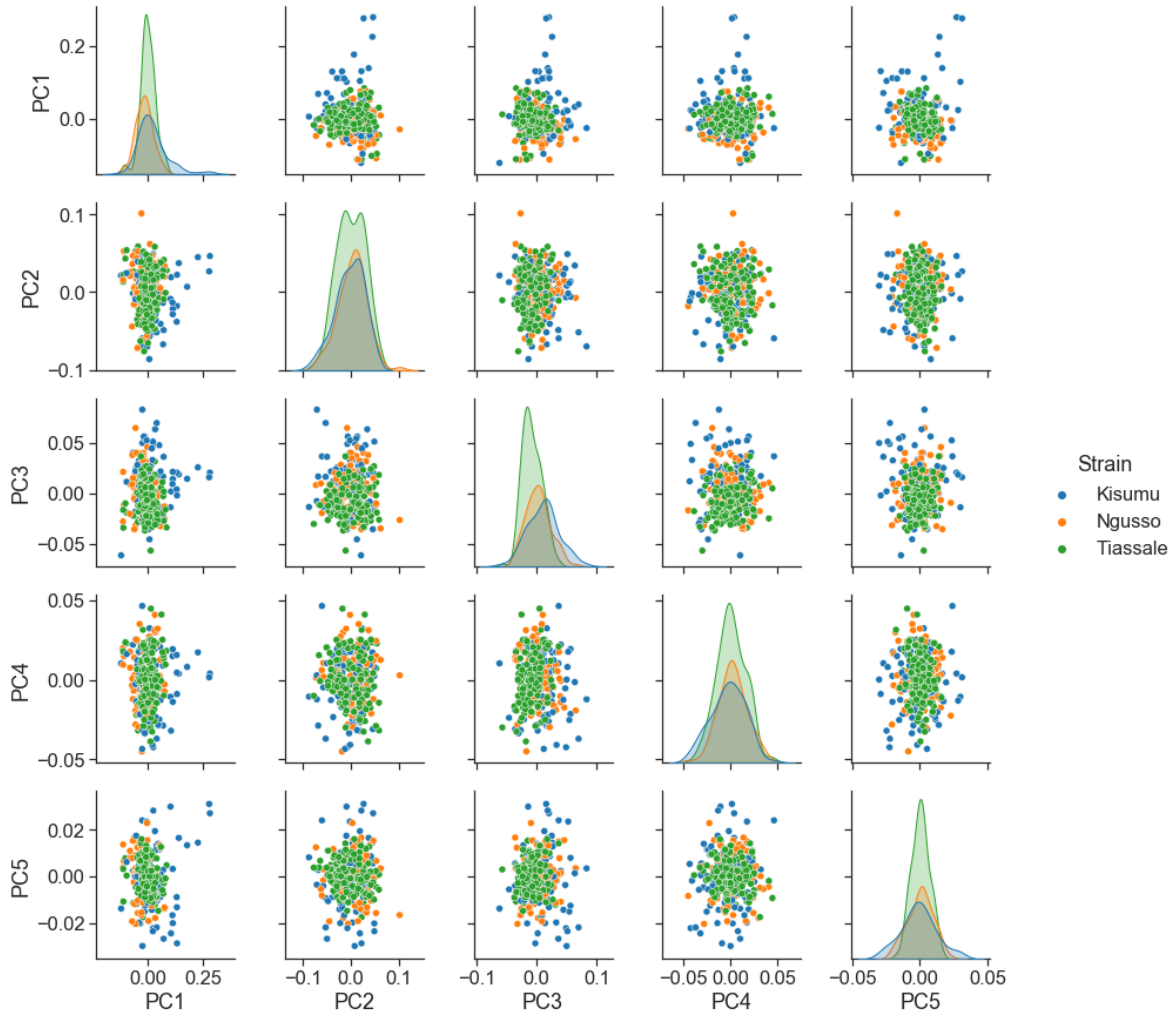


Figure A.9: Pair-plot of PC scores coloured by strain

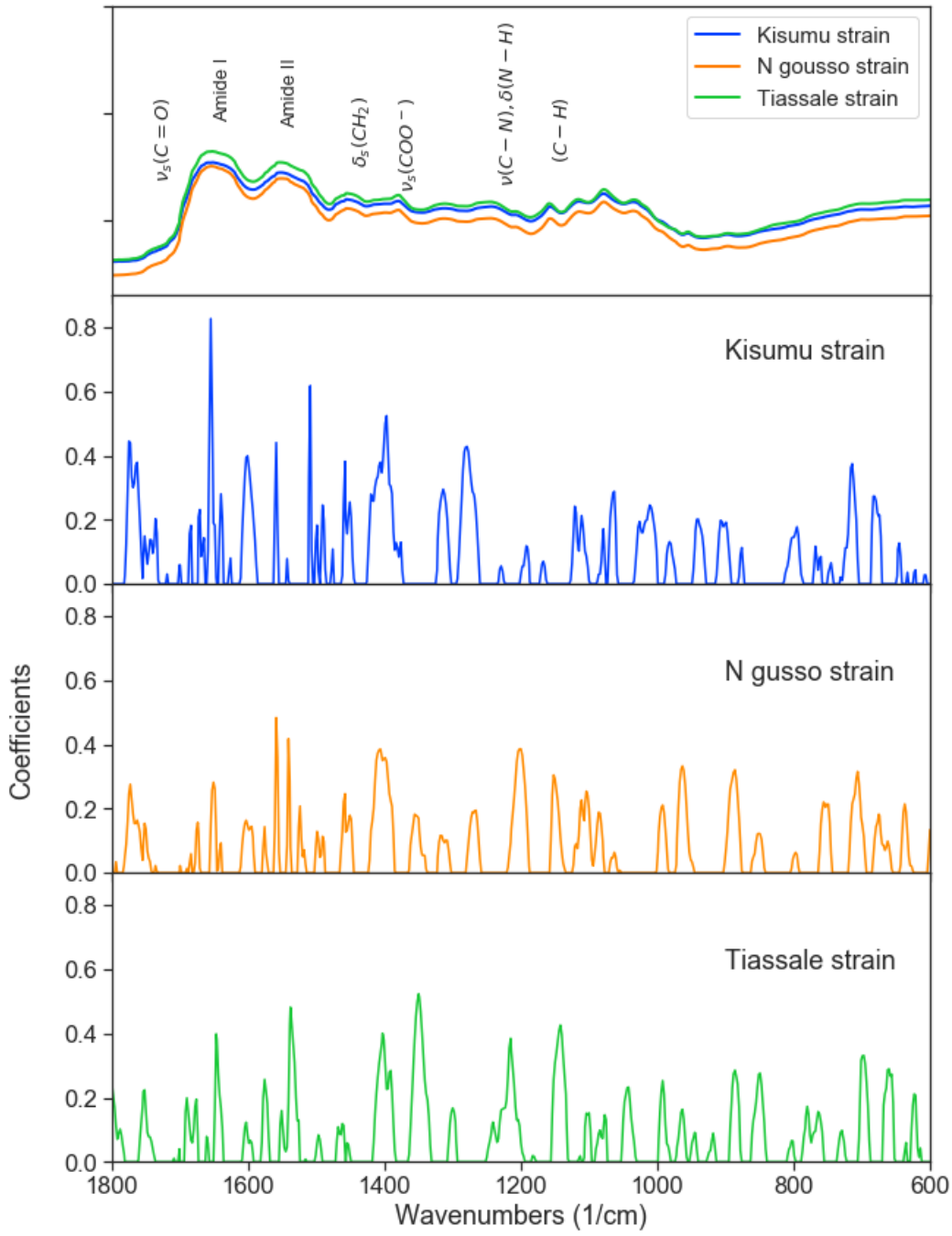


Figure A.10: Prediction coefficients of the optimised model plotted against wavenumbers

Table A.1: Top 10 prediction coefficients for Kisumu, Tiassale and Ngousso strains with tentative functional group assignment

Species Strain	Wavenumber (cm-1)	Functional group Assignment	Aetiological compound
	1397	C-H	Chitin, wax
	1399	C-H	Chitin, wax
	1401	C-H, C-CH3 bend	Chitin, wax
	1509	Amide II	Chitin protein
An. gambiae	1559	Amide II	Chitin protein
Kisumu	1653	Amide I, ν(C=O)	Protein
	1655	Amide I, ν(C=O)	Chitin protein
	1657	Amide I, ν(C=O)	Chitin/protein
	1773	*	*
	1775	*	*
	1198	C-O	*
	1200	C-O	*
	1202	Amide III, ν(C-N), δ(N-H)	*
An. coluzzii	1204	Amide III, ν(C-N), δ(N-H)	*
Ngousso	1404	C-H, C-CH3	Chitin, wax
	1406	C-H, C-CH3	Chitin, wax
	1408	C-H, C-CH3	Chitin, wax
	1410	C-H, C-CH3	Chitin, wax
	1541	Amide II	Protein, chitin
	1559	Amide II	Protein, chitin
	1142	C-O stretch	Chitin
	1144	C-O stretch	Chitin
	1347	*	Protein
	1349	*	Lipids
An. gambiae	1350	*	Lipids
Tiassale	1352	*	

A.1 Hyper-parameter configurations

Logistic regression for species prediction

- penalty = 'l2'
- C=0.1
- class weight = 1:0.5, 0:0.5
- solver= 'liblinear'

Logistic regression for age prediction

- penalty = 'l2'
- C=0.1
- class weight = 1:0.5, 0:0.5
- solver = 'liblinear'

Logistic regression insecticide resistance

- C=0.1
- penalty = 'l2'
- solver = 'saga'

Appendix B

Chapter 3 Appendix

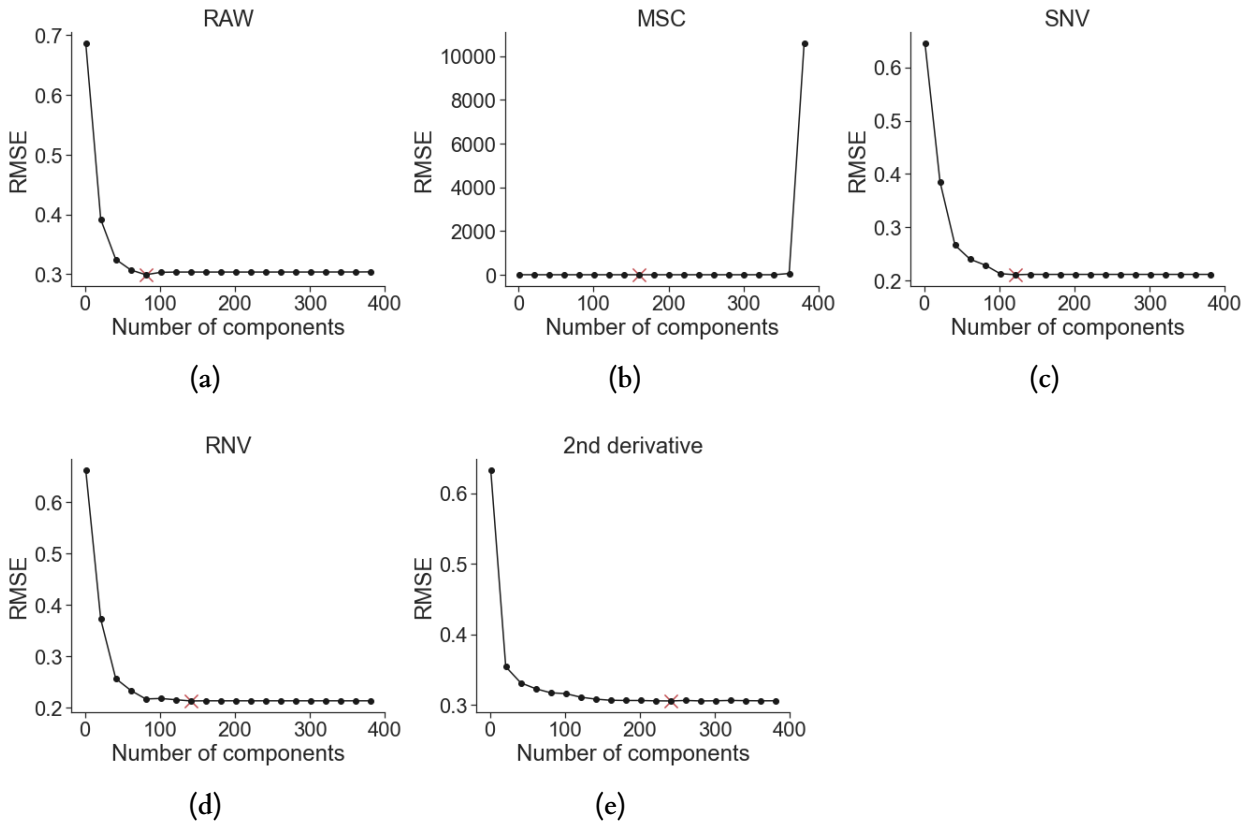


Figure B.1: Selection of optimal number of variables for PLS-DA for species prediction for each pre-processing method. The lowest root mean squared error in cross-validation was selected. The number of component with the lowest RMSE is annotated with a red X. RAW: no pre-processing, MSC: Multiplicative Scatter Correction, SNV: Standard Normal Variate, RNV: Robust Normal Variate. 2nd derivative: Savitzky-Golay filter with 9 point window and second order derivative.

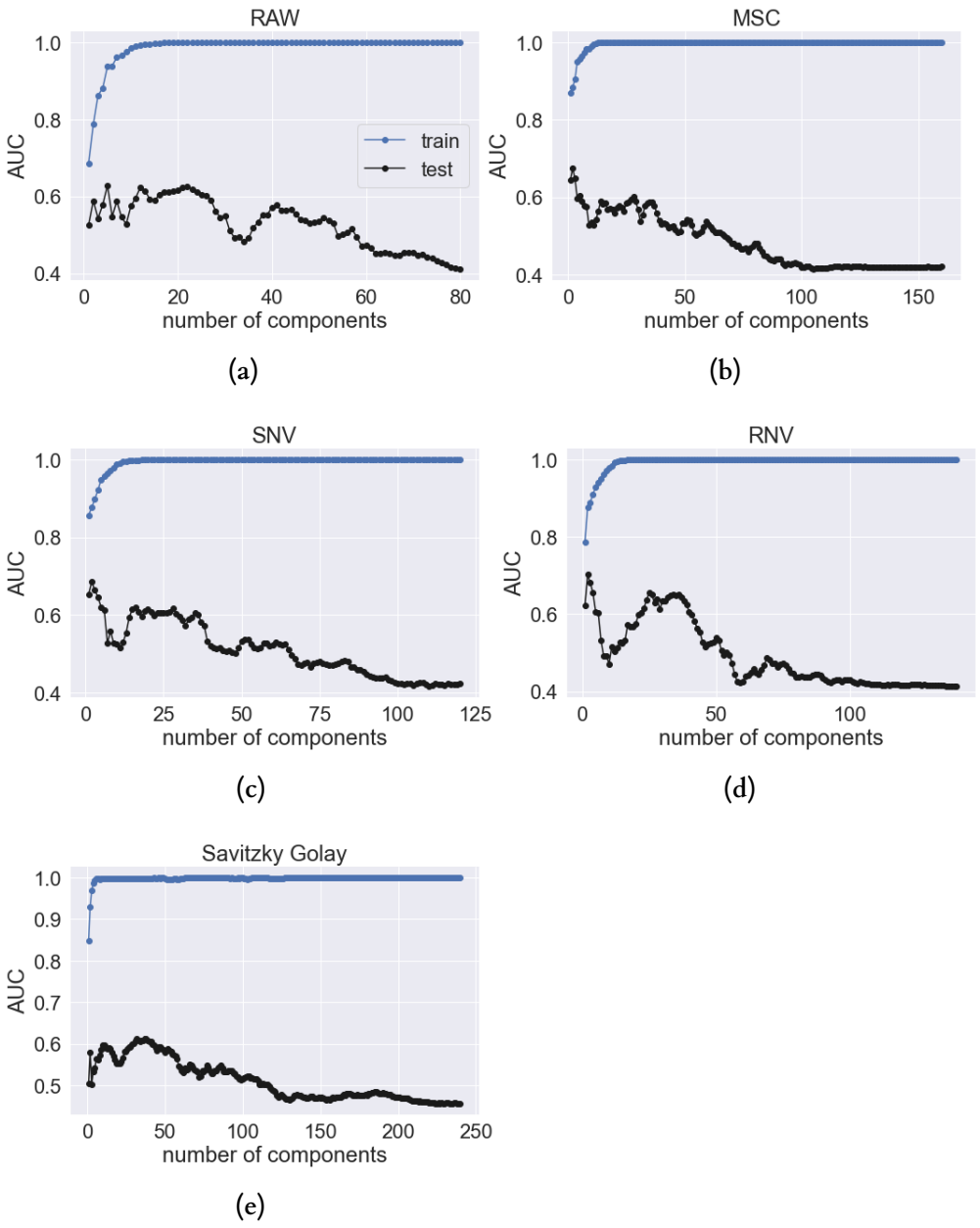


Figure B.2: Scatter plot of AUC values vs number of components when PLS-DA trained in Glasgowlab data set, evaluated with Glasgowlab data set (blue line) and IRSS lab data set (black line) for each pre-processing method. The linear search was set up from 1 to the number of components selected using RMSE.

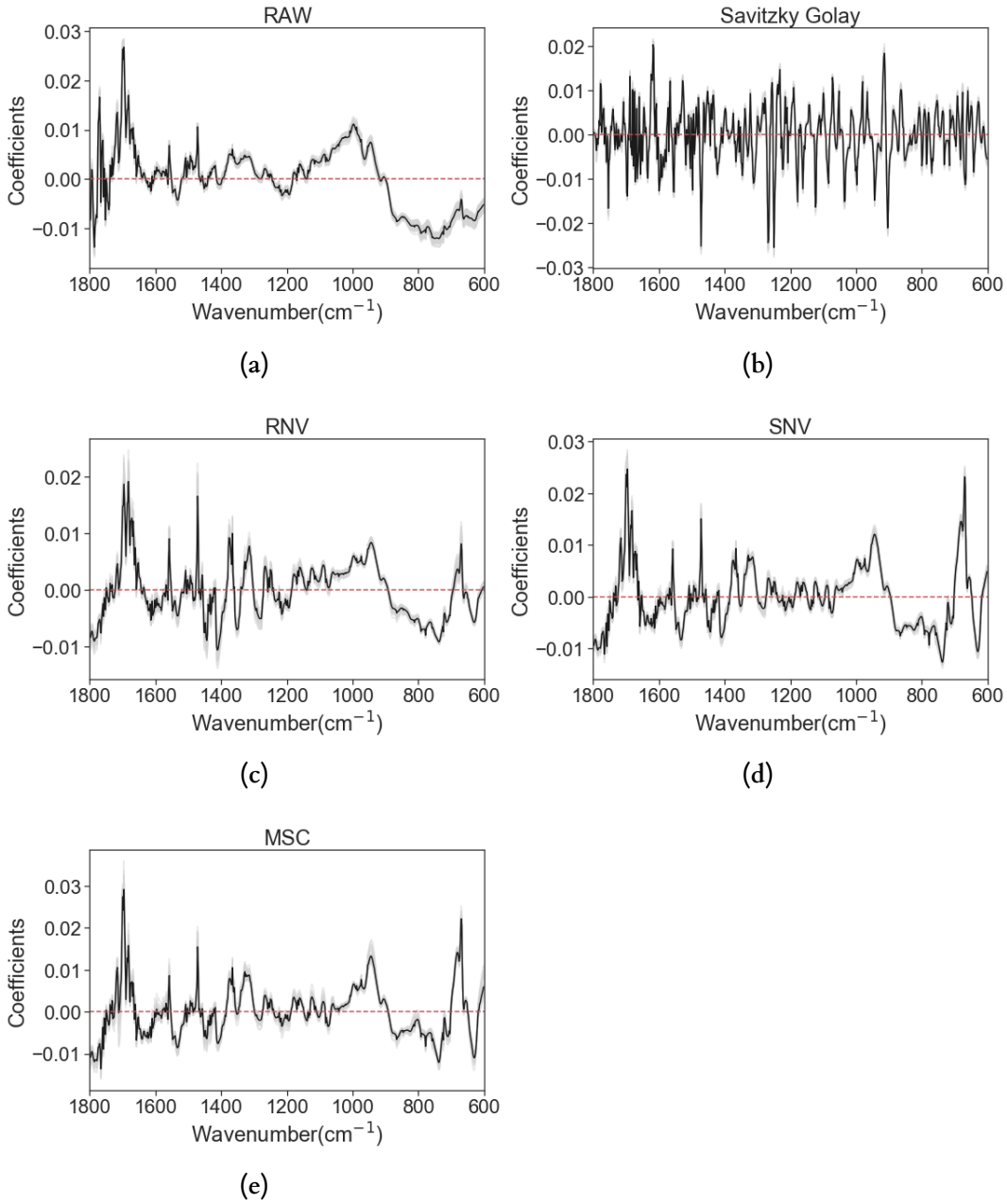


Figure B.3: Model coefficients from calibrated data sets from raw data and after different pre-processing algorithms for species prediction

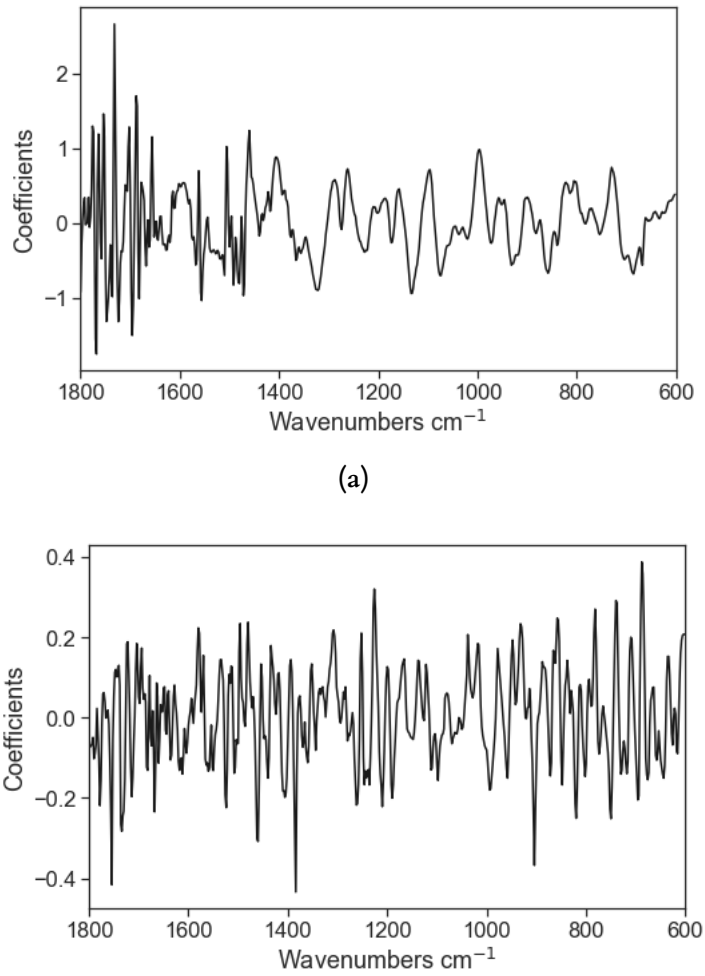


Figure B.4: Model coefficients from calibrated data sets from a) Raw data and after b) Savitzky-Golay preprocessing for age

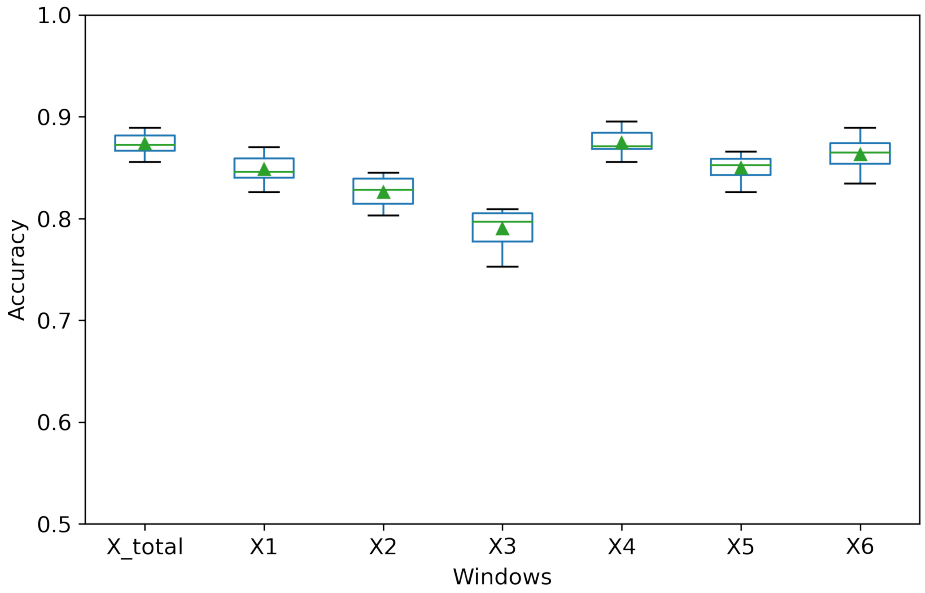


Figure B.6: Boxplot of PLS-DA accuracy using 10-fold cross-validation when using the whole mid infrared spectrum (Total) and each of the spectral windows (X1, X2, X3)

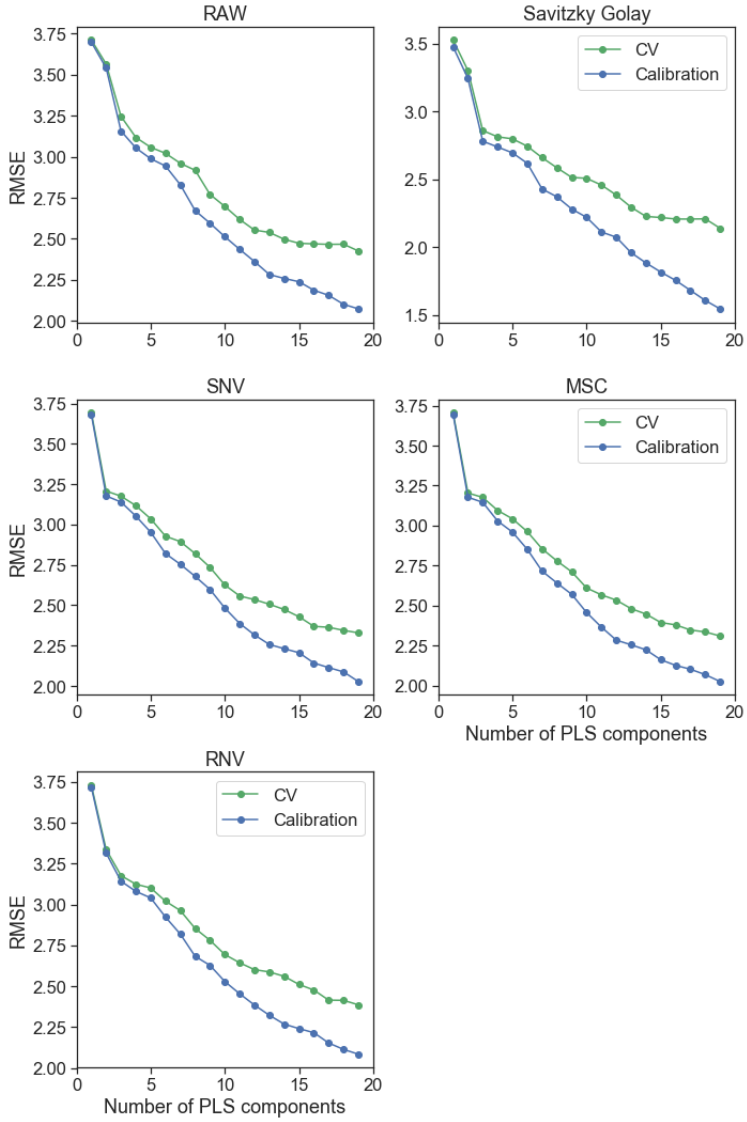


Figure B.5: PLS latent variables optimization for different scattering corrections suing calibration (blue line) and cross-validation (green line) for age prediction

Table B.1: Mean predicted age of *An. gambiae* mosquitoes using Glasgowlab dataset

True age	# samples	Mean predicted Age [95% CI]
1	52	2.21 [2.74 - 1.68]
2	60	3.03[3.48 - 2.59
3	31	3.01 [3.54 - 2.48]
4	46	3.21 [3.83 - 2.59]
5	41	6.44 [7.42 - 5.46]
6	44	7.49 [8.37 - 6.61]
7	42	6.79 [7.93 - 5.65]
8	37	8.28 [8.93 - 7.63]
9	37	9.72 [10.72 - 8.73]
10	47	12.09 [12.77 -11.42]
11	35	11.71 [12.52 - 10.89]
12	45	11.92 [12.65 -11.19]
13	33	12.31[13.28 - 11.34]
14	24	11.29 [12.12 - 10.45]
15	35	13.89 [14.64 - 13.12]
16	38	12.71 [13.26 - 12.15]
17	36	14.21 [15.14 - 13.27]

Table B.2: PLS-DA accuracy for each windows with 10-fold cross-validation (CV), hold out set (Val), and validation with independent sets (IRSSlab and IRSSfield)

Window	CV	Val	IRSSlab	IRSSfield
X total	0.87 ±0.01	0.86	0.60	0.47
X1	0.85 ± 0.02	0.84	0.58	0.45
X2	0.83 ± 0.02	0.82	0.56	0.43
X3	0.79 ± 0.02	0.80	0.61	0.30

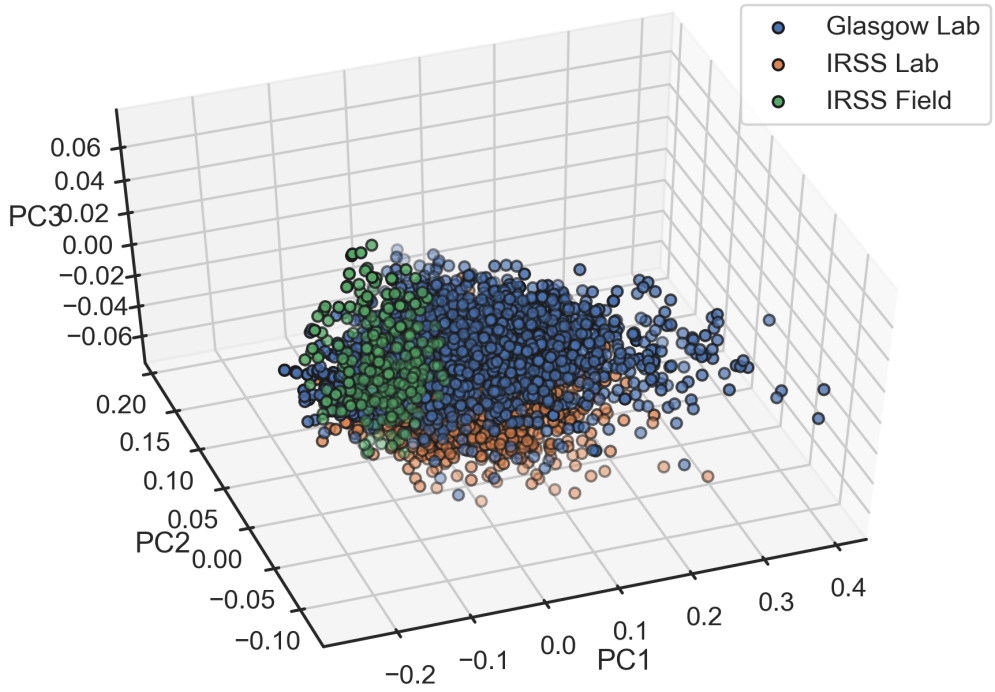


Figure B.7: a) Principal Component Analysis (PCA) PC1 versus PC2 vs PC3 scores plot normalised spectra for laboratory mosquitoes from Glasgowlab (blue) and IRRSlab (orange) and IRSSfield (green)

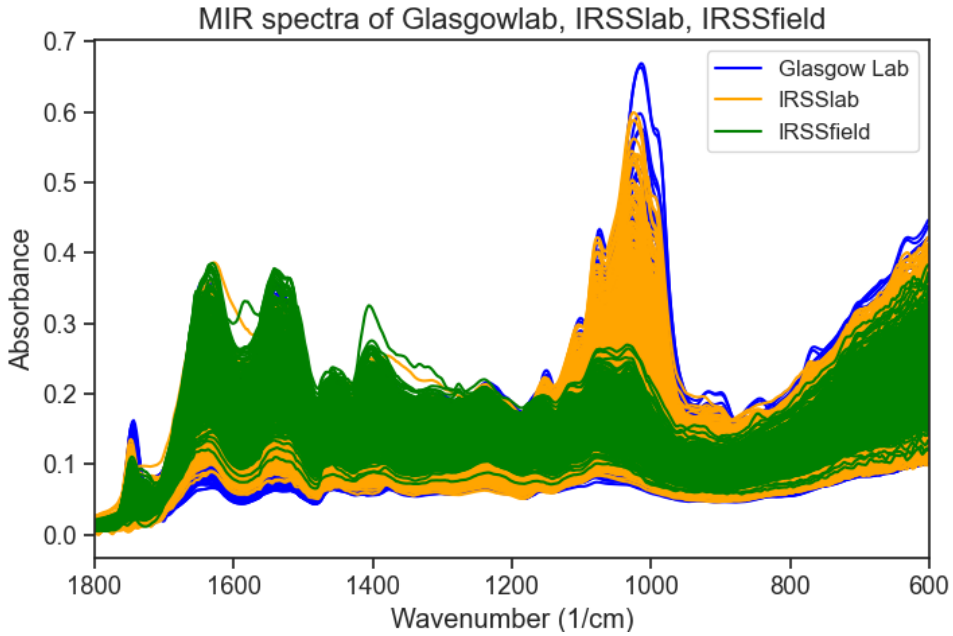


Figure B.8: Mid infrared spectra of all samples from Glasgowlab, IRRSlab and IRSSfield of *An. gambiae* and *An. coluzzii*

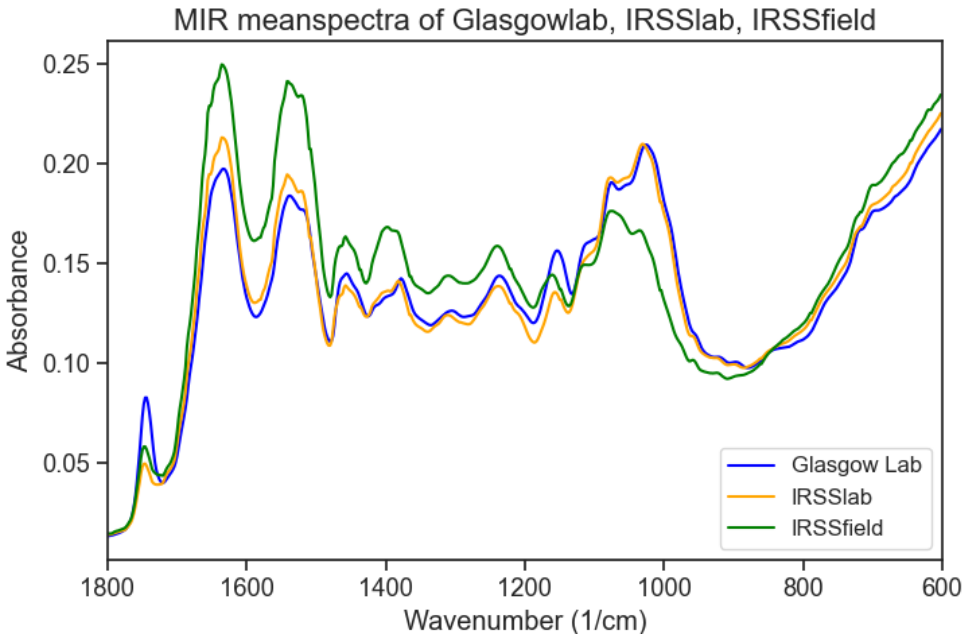


Figure B.9: Mean Mid infrared spectra from Glasgowlab, IRSSlab and IRSSfield

Appendix C

Chapter 4 Appendix

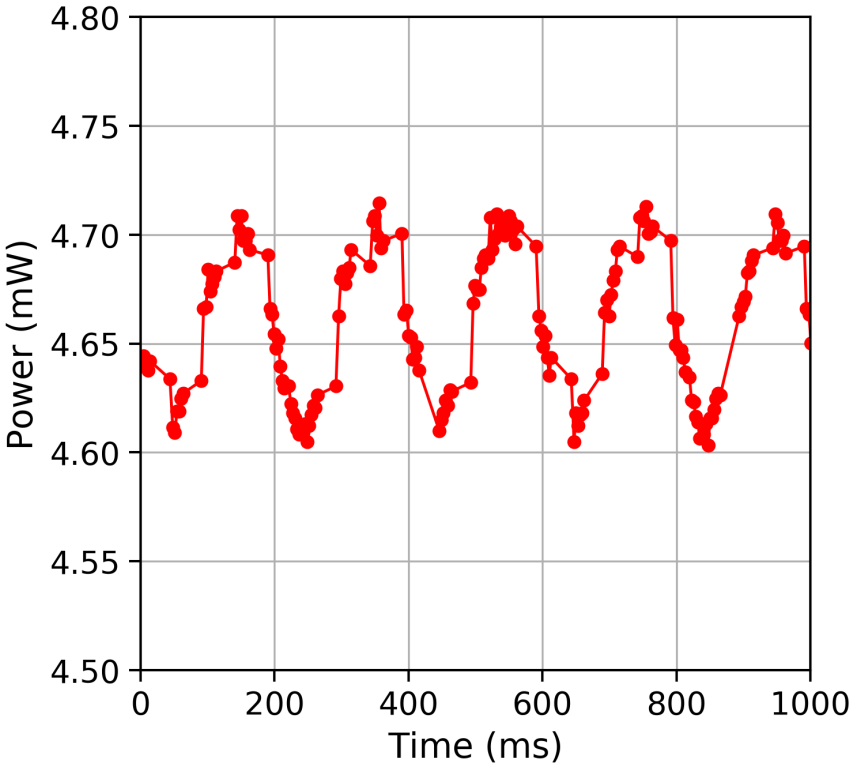


Figure C.1: Power fluctuations of the gain chip in pulse mode (1.45 A, duty cycle 50%, frequency 5 Hz, pulse period 200 ms, pulse width 100 ms) with the galvanometer mirror fixed in one position

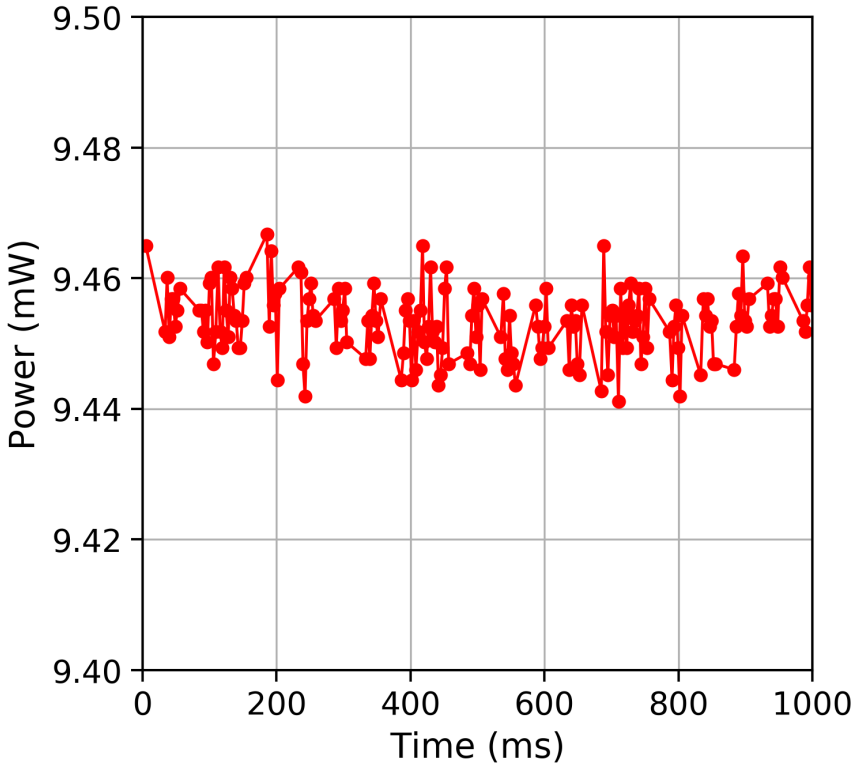


Figure C.2: Power fluctuations of the gain chip in continuous-wave (CW) (1.78 A) with the galvanometer mirror fixed in one position

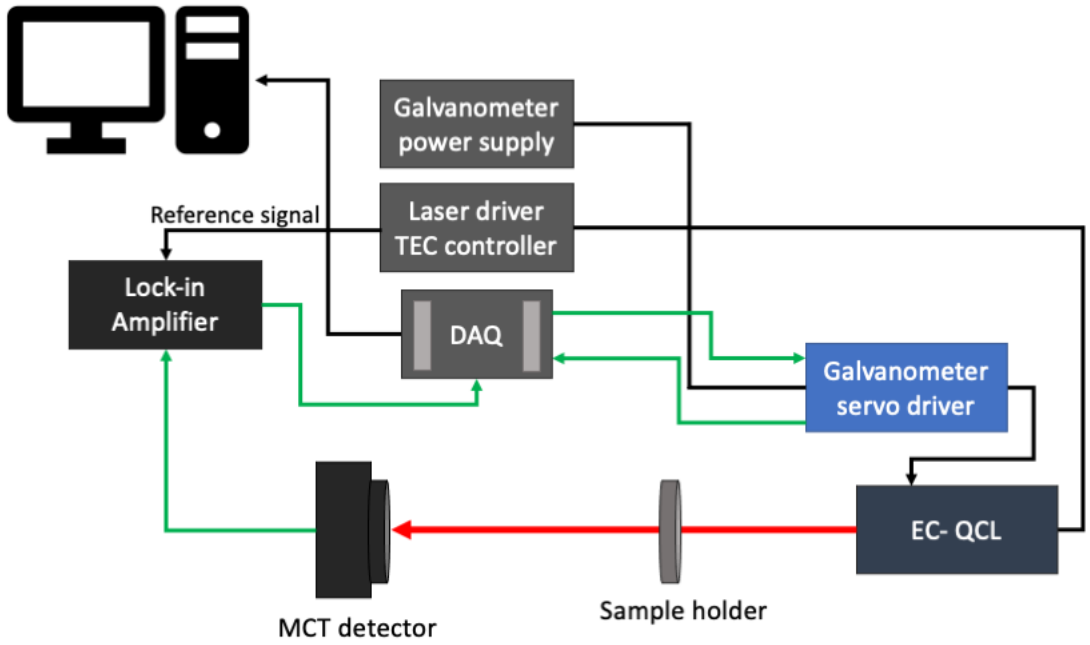


Figure C.3: Schematic of the QCL-based setup mid-IR transmission for solids with the addition of a lock-in amplifier

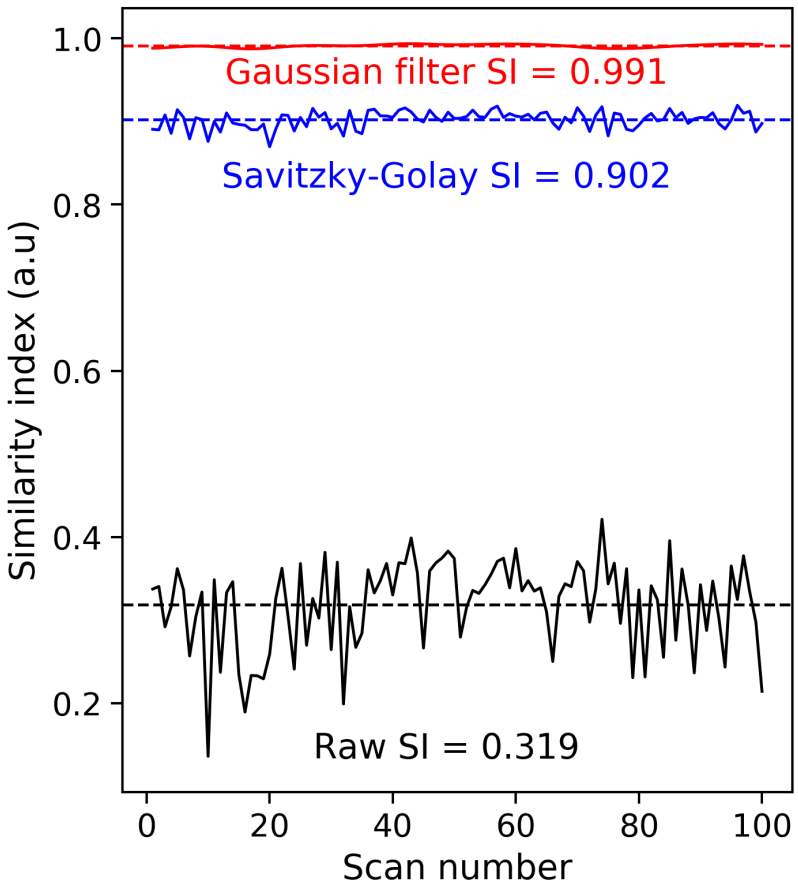


Figure C.4: Comparison of similarity index values between the raw data (black line), after Savitzky-Golay smoothing with a window=21 (blue line), and after Gaussian filter, sigma=0.5 (red line).)

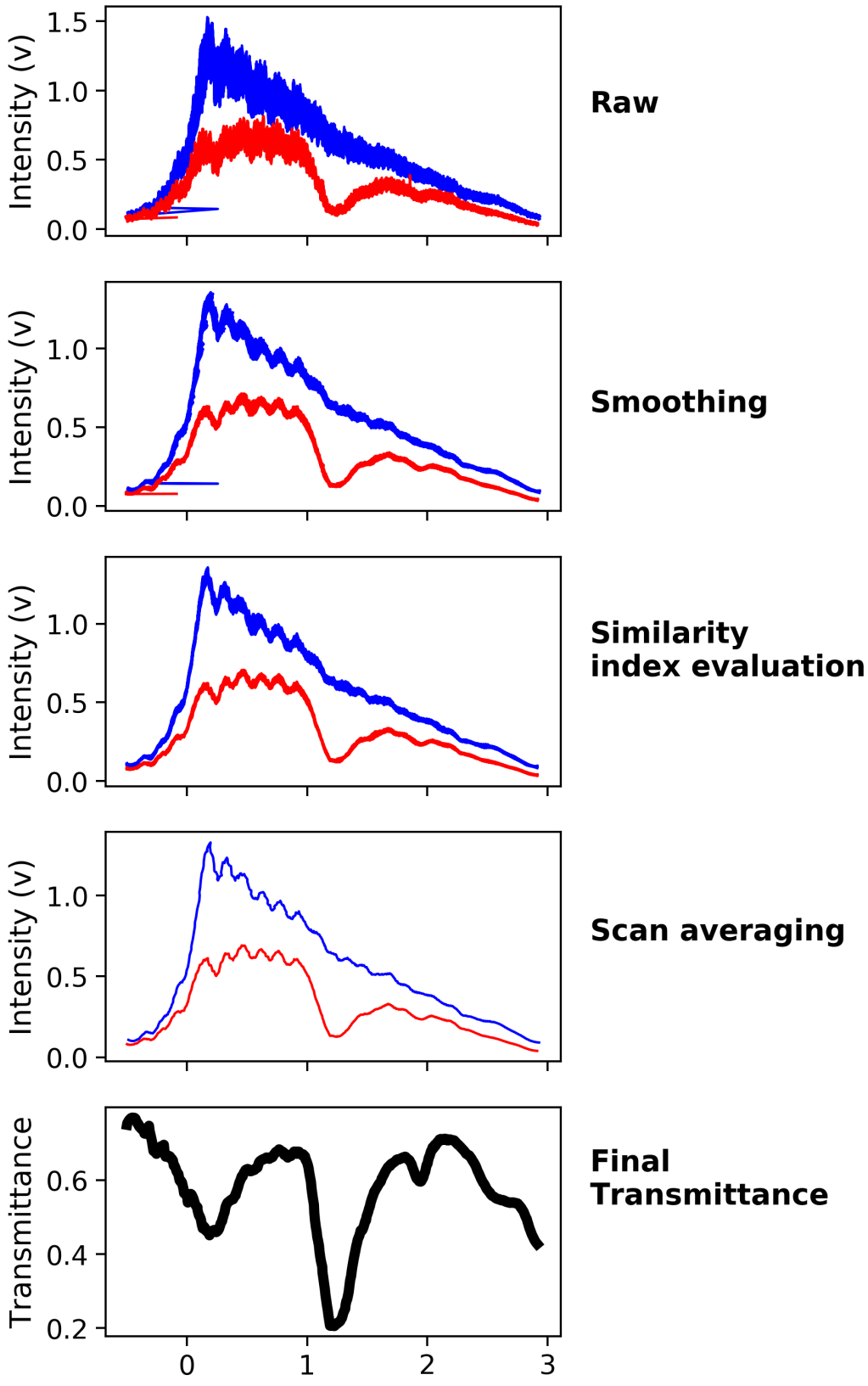


Figure C.5: Processing steps of EC-QCL data using Savitzky-Golay as smoothing filter. Background scans (blue), sample scans (red).

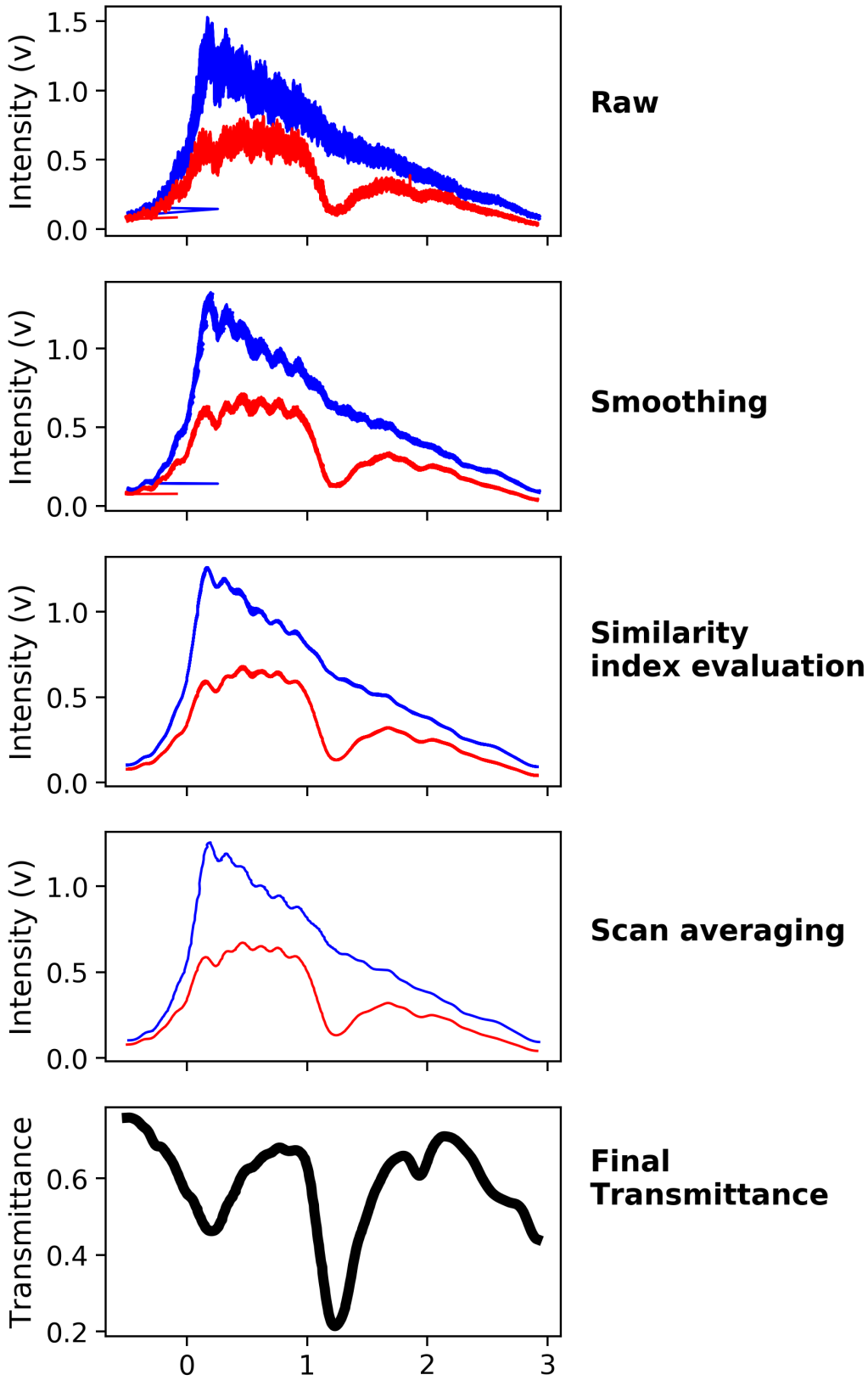


Figure C.6: Processing steps of EC-QCL data using Gaussian filter for signal smoothing. Background scans (blue), sample scans (red).

Bibliography

[1] World Health Organization, "World malaria report 2021," tech. rep., 2021.

[2] M. E. Sinka, M. J. Bangs, S. Manguin, M. Coetzee, C. M. Mbogo, J. Hemingway, A. P. Patil, W. H. Temperley, P. W. Gething, C. W. Kabaria, R. M. Okara, T. Van Boeckel, H. C. J. Godfray, R. E. Harbach, and S. I. Hay, "The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: Occurrence data, distribution maps and bionomic précis," *Parasites and Vectors*, vol. 3, pp. 1–34, dec 2010.

[3] Y. Dahan-Moss, A. Hendershot, M. Dhoogra, H. Julius, J. Zawada, M. Kaiser, N. F. Lobo, B. D. Brooke, and L. L. Koekemoer, "Member species of the Anopheles gambiae complex can be misidentified as Anopheles leesoni," *Malaria Journal*, vol. 19, pp. 1–9, feb 2020.

[4] J. M. Mwangangi, C. M. Mbogo, B. O. Orindi, E. J. Muturi, J. T. Midega, J. Nzovu, H. Gatakaa, J. Githure, C. Borgemeister, J. Keating, and J. C. Beier, "Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years," *Malaria Journal*, vol. 12, pp. 1–9, jan 2013.

[5] Y. A. Afrane, M. Bonizzoni, and G. Yan, *Secondary malaria vectors of sub-Saharan Africa: threat to malaria elimination on the continent?* IntechOpen, 2016.

[6] J. I. Blanford, S. Blanford, R. G. Crane, M. E. Mann, K. P. Paaijmans, K. V. Schreiber, and M. B. Thomas, "Implications of temperature variation for malaria parasite development across Africa," *Scientific Reports*, vol. 3, 2013.

[7] W. R. Shaw and F. Catteruccia, "Vector biology meets disease control: using basic research to fight vector-borne diseases," *Nature Microbiology*, vol. 4, pp. 20–34, jan 2019.

[8] S. Bhatt, D. J. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. E. Battle, C. L. Moyes, A. Henry, P. A. Eckhoff, E. A. Wenger, O. Briët, M. A. Penny, T. A. Smith, A. Bennett, J. Yukich, T. P. Eisele, J. T. Griffin, C. A. Fergus, M. Lynch, F. Lindgren, J. M. Cohen, C. L. Murray, D. L. Smith, S. I. Hay, R. E. Cibulskis, and P. W. Gething, "The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015," *Nature*, vol. 526, pp. 207–211, sep 2015.

[9] O. J. Brady, H. C. J. Godfray, A. J. Tatem, P. W. Gething, J. M. Cohen, F. E. McKenzie, T. A. Perkins, R. C. Reiner, L. S. Tusting, M. E. Sinka, C. L. Moyes, P. A. Eckhoff, T. W. Scott, S. W. Lindsay, S. I. Hay, and D. L. Smith, "Vectorial capacity and vector control: reconsidering sensitivity to parameters for malaria elimination," *Transactions of The Royal Society of Tropical Medicine and Hygiene*, vol. 110, pp. 107–117, feb 2016.

[10] M. Coleman, J. Hemingway, K. A. Gleave, A. Wiebe, P. W. Gething, and C. L. Moyes, "Developing global maps of insecticide resistance risk to improve vector control," *Malaria Journal*, vol. 16, pp. 1–9, feb 2017.

[11] J. Cook, S. Tomlinson, I. Kleinschmidt, M. J. Donnelly, M. Akogbeto, A. Adechoubou, A. Massougbodji, M. Okê-Sopoh, V. Corbel, S. Cornelie, A. Hounto, J. Etang, H. P. Awono-Ambene, J. Bigoga, S. E. Mandeng, B. Njeambosay, R. Tabue, C. Kouambeng, E. Fondjo, K. Raghavendra, R. M. Bhatt, M. K. Chourasia, D. K. Swain, S. Uragayala, N. Valecha, C. Mbogo, N. Bayoh, T. Kinyari, K. Njagi, L. Muthami, L. Kamau, E. Mathenge, E. Ochomo, H. T. Kafy, A. I. Bashir, E. M. Malik, K. Elmardi, J. E. Sulieman, M. Abdin, K. Subramaniam, B. Thomas, P. West, J. Bradley, T. B. Knox, A. P. Mnzava, J. Lines, M. Macdonald, and Z. J. Nkuni, "Implications of insecticide resistance for malaria vector control with long-lasting insecticidal nets: Trends in pyrethroid resistance during a WHO-coordinated multi-country prospective study 11 Medical and Health Sciences 1117 Public Health and Health Se," *Parasites and Vectors*, vol. 11, pp. 1–10, oct 2018.

[12] C. Smith Gueye, G. Newby, R. D. Gosling, M. A. Whittaker, D. Chandramohan, L. Slutsker, and M. Tanner, "Strategies and approaches to vector control in nine malaria-eliminating countries: A cross-case study analysis," *Malaria Journal*, vol. 15, pp. 1–14, jan 2016.

[13] World Health Organization, "WHO guidance note on capacity building in malaria entomology and vector control," tech. rep., Geneva PP - Geneva, 2013.

[14] World Health Organization, "Malaria surveillnace, monitoring & evaluation: A reference manual," tech. rep., 2018.

[15] A. K. Mishra, P. K. Bharti, A. Vishwakarma, S. Nisar, H. Rajvanshi, R. K. Sharma, K. B. Saha, M. M. Shukla, H. Jayswar, A. Das, H. Kaur, S. L. Wattal, and A. A. Lal, "A study of malaria vector surveillance as part of the Malaria Elimination Demonstration Project in Mandla, Madhya Pradesh," *Malaria Journal*, vol. 19, pp. 1–13, dec 2020.

[16] B. J. Johnson, L. E. Hugo, T. S. Churcher, O. T. Ong, and G. J. Devine, "Mosquito Age Grading and Vector-Control Programmes," *Trends in Parasitology*, vol. 36, pp. 39–51, jan 2020.

[17] M. González Jiménez, S. A. Babayan, P. Khazaeli, M. Doyle, F. Walton, E. Reedy, T. Glew, M. Viana, L. Ranford-Cartwright, A. Niang, D. J. Siria, F. O. Okumu, A. Diabaté, H. M.

Ferguson, F. Baldini, and K. Wynne, "Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 3; peer review: 2 approved]," *Wellcome Open Research*, 2019.

[18] V. Mayagaya, A. Ntamatungiro, S. Moore, R. Wirtz, F. Dowell, and M. Maia, "Evaluating preservation methods for identifying Anopheles gambiae s.s. and Anopheles arabiensis complex mosquitoes species using near infra-red spectroscopy," *Parasites & Vectors*, vol. 8, no. 1, p. 60, 2015.

[19] M. Sikulu, K. M. Dowell, L. E. Hugo, R. A. Wirtz, K. Michel, K. H. Peiris, S. Moore, G. F. Killeen, and F. E. Dowell, "Evaluating RNAlater® as a preservative for using near-infrared spectroscopy to predict Anopheles gambiae age and species," *Malaria Journal*, vol. 10, no. 1, p. 186, 2011.

[20] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, vol. 8. 2004.

[21] J. Coates, "a Review of Sampling Methods for Infrared Spectroscopy," in *Applied Spectroscopy* (J. Workman and A. W. Springsteen, eds.), pp. 49–91, San Diego: Academic Press, 1998.

[22] Y. Ozaki, A. A. Christy, and W. F. McClure, *Near-infrared spectroscopy in food science and technology*. John Wiley & Sons, 2006.

[23] P. J. Larkin, *Infrared and Raman Spectroscopy: Principles and Spectral Interpretation*. 2017.

[24] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nature Protocols*, vol. 9, no. 8, pp. 1771–1791, 2014.

[25] J. Haas and B. Mizaikoff, "Advances in Mid-Infrared Spectroscopy for Chemical Analysis," *Annu. Rev. Anal. Chem*, vol. 9, pp. 45–68, 2016.

[26] A. Schwaighofer, M. Brandstetter, and B. Lendl, "Quantum cascade lasers (QCLs) in biomedical spectroscopy," *Chemical Society Reviews*, vol. 46, pp. 5903–5924, oct 2017.

[27] J. M. Hollas, *Basic atomic and molecular spectroscopy*, vol. 11. Royal Society of Chemistry, 2002.

[28] P. W. Atkins, "Physical Chemistry, 6th ed," *Oxford University Press, Oxford*, p. 806, 1998.

[29] J. M. Chalmers and P. R. Griffiths, "Vibrational spectroscopy: sampling techniques and fiber-optics probes," *Applications of vibrational spectroscopy in food science. Wiley, Chichester*, pp. 47–88, 2010.

[30] Y. Ozaki, C. Huck, S. Tsuchikawa, and S. B. Engelsen, *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*. Springer, 2021.

[31] B. Smith, *Infrared spectral interpretation: a systematic approach*. CRC press, 2018.

[32] J. J. Stephanos and A. W. Addison, "Vibrational Rotational Spectroscopy," in *Electrons, Atoms, and Molecules in Inorganic Chemistry*, pp. 505–584, Academic Press, jan 2017.

[33] B. Smith, *Infrared spectral interpretation: a systematic approach*, vol. 36. CRC press, 1999.

[34] N. B. Colthup, "Infrared Spectroscopy," *Encyclopedia of Physical Science and Technology*, pp. 793–816, jan 2003.

[35] A. A. Christy, Y. Ozaki, and V. Gregoriou, "Chapter 5 Group frequencies and assignments of the infrared bands," jan 2001.

[36] H. H. Telle, A. G. Urena, and R. J. Donovan, *Laser chemistry: spectroscopy, dynamics and applications*. John Wiley and Sons, 2007.

[37] J. Workman and A. Springsteen, *Applied Spectroscopy; A compact reference for practitioners*. Academic Press, 1998.

[38] D. W. Ball, "Spectrometer, Spectroscope, and Spectrograph," 2009.

[39] O. Abbas, P. Dardenne, and V. Baeten, "Near-Infrared, Mid-Infrared, and Raman Spectroscopy," in *Chemical Analysis of Food: Techniques and Applications*, pp. 59–89, Academic Press, jan 2012.

[40] V. Saptari, *Fourier-Transform Spectroscopy Instrumentation Engineering | (2003) | Saptari | Publications | Spie*. 2003.

[41] Thermo Fisher Scientific, "Dispersive Infrared Instruments," tech. rep., 2002.

[42] P. J. Larkin, "Instrumentation and Sampling Methods," in *Infrared and Raman Spectroscopy*, pp. 29–61, Elsevier, jan 2018.

[43] M. Hermes, R. B. Morrish, L. Huot, L. Meng, S. Junaid, J. Tomko, G. R. Lloyd, W. T. Masselink, P. Tidemand-Lichtenberg, C. Pedersen, F. Palombo, and N. Stone, "Mid-IR hyperspectral imaging for label-free histopathology and cytology," *Journal of Optics (United Kingdom)*, vol. 20, p. 023002, jan 2018.

[44] J. D. Ellis, "Michelson's Interferometer," in *Field Guide to Displacement Measuring Interferometry*, vol. FG30, pp. 18–19, SPIE, jan 2014.

[45] A. V. Velasco, P. Cheben, M. Florjańczyk, and M. L. Calvo, "Spatial Heterodyne Fourier-Transform Waveguide Spectrometers," in *Progress in Optics*, vol. 59, pp. 159–208, Elsevier, jan 2014.

[46] V. Saptari, *Fourier transform spectroscopy instrumentation engineering*. SPIE Optical Engineering Press Bellingham Washington, DC, 2003.

[47] D. R. Vij, *Handbook of Applied Solid State Spectroscopy*. 2006.

[48] JASCO, "Principles of infrared spectroscopy (3) Principle of FTIR spectroscopy | JASCO Global."

[49] A. Dutta, "Fourier Transform Infrared Spectroscopy," in *Spectroscopic Methods for Nanomaterials Characterization*, vol. 2, pp. 73–93, Elsevier, jan 2017.

[50] D. T. D. Childs, A. B. Krysa, K. L. Kennedy, D. G. Revin, J. W. Cockburn, R. A. Hogg, and S. J. Matcher, "A rapid swept-source mid-infrared laser," in *Conference Digest - IEEE International Semiconductor Laser Conference*, pp. 155–156, IEEE, sep 2014.

[51] M. R. Alcaráz, A. Schwaighofer, C. Kristament, G. Ramer, M. Brandstetter, H. Goicoechea, and B. Lendl, "External-Cavity Quantum Cascade Laser Spectroscopy for Mid-IR Transmission Measurements of Proteins in Aqueous Solution," *Analytical Chemistry*, vol. 87, pp. 6980–6987, jul 2015.

[52] S. Delbeck and H. M. Heise, "FT-IR versus EC-QCL spectroscopy for biopharmaceutical quality assessment with focus on insulin—total protein assay and secondary structure analysis using attenuated total reflection," *Analytical and Bioanalytical Chemistry*, vol. 412, pp. 4647–4658, jun 2020.

[53] C. K. Akhgar, G. Ramer, M. Żbik, A. Trajnerowicz, J. Pawluczyk, A. Schwaighofer, and B. Lendl, "The Next Generation of IR Spectroscopy: EC-QCL-Based Mid-IR Transmission Spectroscopy of Proteins with Balanced Detection," *Analytical Chemistry*, vol. 92, pp. 9901–9907, jul 2020.

[54] K. Isensee, N. Kröger-Lui, and W. Petrich, "Biomedical Applications of Mid-Infrared Quantum Cascade Lasers – a Review," *The Analyst*, 2018.

[55] W. Perkins, "Sample handling in infrared spectroscopy–an overview," *Practical Sampling Techniques for Infrared Analysis*, pp. 11–53, 1993.

[56] Q. Ye and P. Spencer, "Analyses of material-tissue interfaces by Fourier transform infrared, Raman spectroscopy, and chemometrics," in *Material-Tissue Interfacial Phenomena: Contributions from Dental and Craniofacial Reconstructions*, pp. 231–251, Woodhead Publishing, jan 2017.

[57] N. Abidi, "Fourier transform infrared spectroscopy: Developments, techniques and applications," in *Fourier Transform Infrared Spectroscopy: Developments, Techniques and Applications*, pp. 139–158, Nova Science Publishers, Inc., 2011.

[58] M. P. Fuller and P. R. Griffiths, "Diffuse reflectance measurements by infrared fourier transform spectrometry," *Analytical chemistry*, vol. 50, no. 13, pp. 1906–1910, 1978.

[59] M. Tasumi, *Introduction to experimental infrared spectroscopy: Fundamentals and practical methods*. Wyley and Sons, 2014.

[60] J. Katon, "Infrared microspectroscopy. a review of fundamentals and applications," *Micron*, vol. 27, no. 5, pp. 303–314, 1996.

[61] H. J. Humecki, *Practical guide to infrared microspectroscopy*. CRC Press, 1995.

[62] I. Reiche and E. Chalmin, "Synchrotron Methods: Color in Paints and Minerals," in *Treatise on Geochemistry: Second Edition*, vol. 14, pp. 209–239, Elsevier, jan 2013.

[63] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science*, vol. 264, no. 5158, pp. 553–556, 1994.

[64] Y. Yao, A. J. Hoffman, and C. F. Gmachl, "Mid-infrared quantum cascade lasers," *Nature Photonics*, vol. 6, pp. 432–439, jul 2012.

[65] R. F. Kazarinov and R. A. Suris., "Possible amplification of electromagnetic waves in a semiconductor with a superlattice," *Fiz. Tekh. Poluprovodn.*, vol. 5, no. 4, pp. 707–709, 1971.

[66] P. I. Abramov, E. V. Kuznetsov, L. A. Skvortsov, and M. I. Skvortsova, "Quantum-Cascade Lasers in Medicine and Biology (Review)," *Journal of Applied Spectroscopy*, vol. 86, no. 1, pp. 1–26, 2019.

[67] R. Kasahara, S. Kino, S. Soyama, and Y. Matsuura, "Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers," *Biomedical Optics Express*, vol. 9, p. 289, jan 2018.

[68] O. Spitz, A. Herdt, J. Wu, G. Maisons, M. Carras, C. W. Wong, W. Elsäßer, and F. Grillot, "Private communication with quantum cascade laser photonic chaos," *Nature Communications*, vol. 12, pp. 1–8, jun 2021.

[69] J. Wagner, R. Ostendorf, J. Grahmann, A. Merten, S. Hugger, J.-P. Jarvis, F. Fuchs, D. Boskovic, and H. Schenk, "Widely tunable quantum cascade lasers for spectroscopic sensing," *Proc. SPIE*, vol. 9370, p. 937012, feb 2015.

[70] O. Malis, C. Gmachl, D. L. Sivco, L. N. Pfeiffer, A. Michael Sergent, and K. W. West, "The quantum cascade laser: A versatile high-power semiconductor laser for mid-infrared applications," *Bell System Technical Journal*, vol. 10, no. 3, pp. 199–214, 2005.

[71] J. Faist, *Quantum cascade lasers*. OUP Oxford, 2013.

[72] R. Pecharroman-Gallego, "An Overview on Quantum Cascade Lasers: Origins and Development," in *Quantum Cascade Lasers*, InTech, apr 2017.

[73] M. A. Belkin and F. Capasso, "New frontiers in quantum cascade lasers: High performance room temperature terahertz sources," *Physica Scripta*, vol. 90, no. 11, 2015.

[74] A. Schwaighofer and B. Lendl, "Quantum cascade laser-based infrared transmission spectroscopy of proteins in solution," in *Vibrational Spectroscopy in Protein Research*, pp. 59–88, Elsevier, jan 2020.

[75] F. Capasso, "High-performance midinfrared quantum cascade lasers," *SPIE Reviews*, vol. 1, p. 111102, nov 2010.

[76] F. Bassani, G. L. Liedl, and P. Wyder, *Encyclopedia of condensed matter physics*. Elsevier, 2005.

[77] C. Sirtori and R. Teissier, "Quantum Cascade Lasers: Overview of Basic Principles of Operation and State of the Art," in *Intersubband Transitions In Quantum Structures.*, ch. Quantum Ca, pp. 1–44, Oxford University Press, mar 2006.

[78] R. Colombelli, F. Capasso, C. Gmachl, A. L. Hutchinson, D. L. Sivco, A. Tredicucci, M. C. Wanke, A. M. Sergent, and A. Y. Cho, "Far-infrared surface-plasmon quantum-cascade lasers at 21.5 μm and 24 μm wavelengths," *Applied physics letters*, vol. 78, no. 18, pp. 2620–2622, 2001.

[79] A. Pabjańczyk, R. Sarzała, M. Wasiak, and M. Bugajski, "Kwantowe lasery kaskadowe: podstawy fizyczne," *Elektronika: konstrukcje, technologie, zastosowania*, vol. 50, no. 5, pp. 30–43, 2009.

[80] H. Detz, A. M. Andrews, M. A. Kainz, S. Schönhuber, T. Zederbauer, D. MacFarland, M. Krall, C. Deutsch, M. Brandstetter, P. Klang, W. Schrenk, K. Unterrainer, and G. Strasser, "Evaluation of Material Systems for THz Quantum Cascade Laser Active Regions," *Physica Status Solidi (A) Applications and Materials Science*, vol. 216, p. 1800504, jan 2019.

[81] M. S. Vitiello, G. Scalari, B. Williams, and P. De Natale, "Quantum cascade lasers: 20 years of challenges," *Optics Express*, vol. 23, no. 4, p. 5167, 2015.

[82] J. Faist, C. Gmachl, F. Capasso, C. Sirtori, D. L. Sivco, J. N. Baillargeon, and A. Y. Cho, "Distributed feedback quantum cascade lasers," *Applied Physics Letters*, vol. 70, no. 20, pp. 2670–2672, 1997.

[83] R. Lewicki, M. Witinski, B. Li, and G. Wysocki, "Spectroscopic benzene detection using a broadband monolithic dfb-qcl array," in *Novel In-Plane Semiconductor Lasers XV*, vol. 9767, p. 97671T, International Society for Optics and Photonics, 2016.

[84] G. N. Rao and A. Karpf, "External cavity tunable quantum cascade lasers and their applications to trace gas monitoring," *Applied Optics*, vol. 50, p. A100, feb 2011.

[85] G. P. Luo, C. Peng, H. Q. Le, S. S. Pei, W. Y. Hwang, B. Ishaug, J. Um, J. N. Baillargeon, and C. H. Lin, "Grating-tuned external-cavity quantum-cascade semiconductor lasers," *Applied Physics Letters*, vol. 78, no. 19, pp. 2834–2836, 2001.

[86] B. Meng and Q. J. Wang, "Broadly tunable single-mode mid-infrared quantum cascade lasers," *Journal of Optics (United Kingdom)*, vol. 17, p. 023001, jan 2015.

[87] A. Hugi, R. Terazzi, Y. Bonetti, A. Wittmann, M. Fischer, M. Beck, J. Faist, and E. Gini, "External cavity quantum cascade laser tunable from 7.6 to 11.4 μm," *Applied Physics Letters*, vol. 95, p. 061103, aug 2009.

[88] C. W. Liu, J. C. Zhang, F. L. Yan, Z. W. Jia, Z. B. Zhao, N. Zhuo, F. Q. Liu, and Z. G. Wang, "External Cavity Tuning of Coherent Quantum Cascade Laser Array Emitting at 7.6 μm," *Chinese Physics Letters*, vol. 34, p. 034209, mar 2017.

[89] K. Liu and M. G. Littman, "Novel geometry for single-mode scanning of tunable lasers," *Optics Letters*, vol. 6, p. 117, mar 1981.

[90] L. Butschek, S. Hugger, and J. Jarvis, "Microoptoelectromechanical systems-based external cavity quantum cascade lasers for real-time spectroscopy," *Optical Engineering*, vol. 57, p. 1, sep 2017.

[91] C. K. N. Patel, "Advances in Fabry-Perot and tunable quantum cascade lasers," vol. 10194, p. 101942H, International Society for Optics and Photonics, may 2017.

[92] F. K. Tittel and R. Lewicki, "Tunable mid-infrared laser absorption spectroscopy," in *Semiconductor Lasers: Fundamentals and Applications*, pp. 579–629, Woodhead Publishing, jan 2013.

[93] G. D. Banik, S. Som, A. Maity, M. Pal, S. Maithani, S. Mandal, and M. Pradhan, "An EC-QCL based N 2 O sensor at 5.2 μm using cavity ring-down spectroscopy for environmental applications," *Analytical Methods*, vol. 9, pp. 2315–2320, apr 2017.

[94] B.-U. Sohn, P. Xing, and D. T. H. Tan, "Compact open-path detection of N 2 O gas with low concentration of ppb level based on QCL," *2017 Conference on Lasers and Electro-Optics Pacific Rim (CLEO-PR)*, pp. 2–3, jul 2017.

[95] W. Ren, W. Jiang, and F. K. Tittel, "Single-QCL-based absorption sensor for simultaneous trace-gas detection of CH4 and N2O," *Applied Physics B: Lasers and Optics*, vol. 117, pp. 245–251, oct 2014.

[96] C. J. Breshike, C. A. Kendziora, R. Furstenberg, R. A. McGill, A. Kusterbeck, and V. Nguyen, "Using infrared backscatter imaging spectroscopy to detect trace explosives at standoff distances," in *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XIX* (A. W. Fountain, J. A. Guicheteau, and C. R. Howle, eds.), vol. 10629, p. 22, SPIE, may 2018.

[97] J. Miles, M. F. Muir, S. W. Scully, I. Hunter, C. McIlroy, J. E. Cunningham, C. V. Robinson, and C. R. Howle, "Towards standoff photothermal spectroscopy of CBRNE hazards using an ultra-fast tunable QCL," vol. 11416, p. 18, SPIE-Intl Soc Optical Eng, apr 2020.

[98] A. K. Goyal, D. B. Kelley, D. A. Wood, and P. Kotidis, "High-speed and large-area scanning of surfaces for trace chemicals using wavelength-tunable quantum cascade lasers," in *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XIX* (A. W. Fountain, J. A. Guicheteau, and C. R. Howle, eds.), vol. 10629, p. 8, SPIE, may 2018.

[99] E. Goormaghtigh, "Infrared imaging in histopathology: Is a unified approach possible?," *Biomedical Spectroscopy and Imaging*, vol. 5, no. 4, pp. 325–346, 2017.

[100] S. Mittal and R. Bhargava, "A comparison of mid-infrared spectral regions on accuracy of tissue classification," *Analyst*, vol. 144, pp. 2635–2642, apr 2019.

[101] Y.-P. Tseng, P. Bouzy, C. Pedersen, N. Stone, and P. Tidemand-Lichtenberg, "Upconversion raster scanning microscope for long-wavelength infrared imaging of breast cancer microcalcifications," *Biomedical Optics Express*, vol. 9, p. 4979, oct 2018.

[102] M. J. Pilling, A. Henderson, and P. Gardner, "Quantum Cascade Laser Spectral Histopathology: Breast Cancer Diagnostics Using High Throughput Chemical Imaging," *Analytical Chemistry*, vol. 89, pp. 7348–7355, jul 2017.

[103] C. Kuepper, A. Kallenb, H. Juette, A. Tannapfel, F. Großerueschkamp, K. Gerwert, A. Kallenbach-Thieltges, H. Juette, A. Tannapfel, F. Großerueschkamp, and K. Gerwert, "Quantum Cascade Laser-Based Infrared Microscopy for Label-Free and Automated Cancer Classification in Tissue Sections," *Scientific Reports*, vol. 8, no. 1, p. 7717, 2018.

[104] D. Liberda, M. Hermes, P. Koziol, N. Stone, and T. P. Wrobel, "Translation of an esophagus histopathological FT-IR imaging model to a fast quantum cascade laser modality," *Journal of Biophotonics*, vol. 13, aug 2020.

[105] M. Pleitez, H. Von Lilienfeld-Toal, and W. Mäntele, "Infrared spectroscopic analysis of human interstitial fluid in vitro and in vivo using FT-IR spectroscopy and pulsed quantum cascade lasers (QCL): Establishing a new approach to non invasive glucose measurement," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 85, pp. 61–65, jan 2012.

[106] M. Grafen, K. Nalpantidis, A. Ostendorf, D. Ihrig, and H. M. Heise, "Characterization of a multi-module tunable EC-QCL system for mid-infrared biofluid spectroscopy for hospital use and personalized diabetes technology," in *Optical Diagnostics and Sensing XVI: Toward Point-of-Care Diagnostics* (G. L. Coté, ed.), vol. 9715, p. 97150T, International Society for Optics and Photonics, mar 2016.

[107] T. Koyama, S. Kino, and Y. Matsuura, "Accuracy Improvement of Blood Glucose Measurement System Using Quantum Cascade Lasers," *Optics and Photonics Journal*, vol. 09, no. 10, pp. 155–164, 2019.

[108] A. Ostendorf, M. Grafen, S. Delbeck, H. Busch, and H. M. Heise, "Evaluation and benchmarking of an EC-QCL-based mid-infrared spectrometer for monitoring metabolic blood parameters in critical care units," in *Proc. SPIE 10501, Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics, 105010A*, vol. 10501, p. 9, SPIE, feb 2018.

[109] I. L. Jernelv, K. Strøm, D. R. Hjelme, and A. Aksnes, "Mid-infrared spectroscopy with a fiber-coupled tuneable quantum cascade laser for glucose sensing," in *Optical Fibers and Sensors for Medical Diagnostics and Treatment Applications XX* (I. Gannot, ed.), vol. 11233, p. 36, SPIE, feb 2020.

[110] A. Werth, S. Liakat, A. Dong, C. M. Woods, and C. F. Gmachl, "Implementation of an integrating sphere for the enhancement of noninvasive glucose detection using quantum cascade laser spectroscopy," *Applied Physics B: Lasers and Optics*, vol. 124, p. 75, may 2018.

[111] M. A. Pleitez, O. Hertzberg, A. Bauer, M. Seeger, T. Lieblein, H. V. Lilienfeld-Toal, and W. Mäntele, "Photothermal deflectometry enhanced by total internal reflection enables non-invasive glucose monitoring in human epidermis," *Analyst*, vol. 140, pp. 483–488, jan 2015.

[112] Y. Matsuura and T. Koyama, "Non-invasive blood glucose measurement using quantum cascade lasers," in *Quantum Sensing and Nano Electronics and Photonics XVI* (M. Razeghi, J. S. Lewis, G. A. Khodaparast, and E. Tournié, eds.), p. 6, SPIE, feb 2019.

[113] R. Centeno, J. Mandon, F. J. Harren, and S. M. Cristescu, "Influence of ethanol on breath acetone measurements using an external cavity quantum cascade laser," *Photonics*, vol. 3, p. 22, apr 2016.

[114] R. Ghorbani and F. M. Schmidt, "Real-time breath gas analysis of CO and CO2 using an EC-QCL," *Applied Physics B: Lasers and Optics*, vol. 123, p. 144, may 2017.

[115] H. Tian, L. Cao, D. An, T. He, and J. Li, "Quantum cascade laser spectroscopic sensor for breath gas analysis," in *2017 International Conference on Optical Instruments and Technology: Advanced Optical Sensors and Applications* (L. Dong, X. Zhang, H. Xiao, and F. J. Arregui, eds.), vol. 10618, p. 27, SPIE, jan 2018.

[116] A. P. M. Michel, A. E. Morrison, B. C. Colson, W. A. Pardis, X. A. Moya, C. C. Harb, and H. K. White, "Quantum cascade laser-based reflectance spectroscopy: a robust approach for the classification of plastic type," *Optics Express*, vol. 28, no. 12, p. 17741, 2020.

[117] S. Primpke, M. Godejohann, and G. Gerdts, "Rapid Identification and Quantification of Microplastics in the Environment by Quantum Cascade Laser-Based Hyperspectral Infrared Chemical Imaging," *Environmental Science and Technology*, vol. 54, pp. 15893–15903, dec 2020.

[118] P. Torrione, L. M. Collins, and K. D. Morton, "Multivariate analysis, chemometrics, and machine learning in laser spectroscopy," in *Laser Spectroscopy for Sensing: Fundamentals, Techniques and Applications*, pp. 125–164, Woodhead Publishing, jan 2014.

[119] K. M. Mendez, D. I. Broadhurst, and S. N. Reinke, "Migrating from partial least squares discriminant analysis to artificial neural networks: a comparison of functionally equivalent visualisation and feature contribution tools using jupyter notebooks," *Metabolomics*, vol. 16, p. 17, feb 2020.

[120] M. Paluszek, S. Thomas, M. Paluszek, and S. Thomas, "An Overview of Machine Learning," in *MATLAB Machine Learning Recipes*, pp. 1–18, Apress, 2019.

[121] P. Puthongkham, S. Wirojsaengthong, and A. Suea-Ngam, "Machine learning and chemometrics for electrochemical sensors: moving forward to the future of analytical chemistry," *The Analyst*, vol. 146, pp. 6351–6364, oct 2021.

[122] A. Panesar, *What Is Machine Learning?*, pp. 75–118. Berkeley, CA: Apress, 2019.

[123] I. El Naqa and M. J. Murphy, "What Is Machine Learning? BT - Machine Learning in Radiation Oncology: Theory and Applications," pp. 3–11, Cham: Springer International Publishing, 2015.

[124] A. Geron, *Hands–On Machine Learning with Scikit–Learn and TensorFlow 2nd edition*. O'Reilly Media, Inc., 2019.

[125] T. J. Cleophas and A. H. Zwinderman, *Machine Learning in Medicine-a Complete Overview*. Springer, 2015.

[126] O. V. Prezhdo, "Advancing physical chemistry with machine learning," 2020.

[127] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[128] A. Fielding, *Machine learning methods for ecological applications*. Springer Science & Business Media, 1999.

[129] T. P. Trappenberg, *Fundamentals of Machine Learning*. Oxford University Press, nov 2019.

[130] J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[131] L. B. Ayres, F. J. Gomez, J. R. Linton, M. F. Silva, and C. D. Garcia, "Taking the leap between analytical chemistry and artificial intelligence: A tutorial review," may 2021.

[132] F. C. Pereira and S. S. Borysov, "Machine learning fundamentals," in *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling*, pp. 9–29, Elsevier, jan 2018.

[133] J. E. Crawford, M. M. Riehle, K. Markianos, E. Bischoff, W. M. Guelbeogo, A. Gneme, N. Sagnon, K. D. Vernick, R. Nielsen, and B. P. Lazzaro, "Evolution of GOUNDRY, a cryptic subgroup of Anopheles gambiae s.l., and its impact on susceptibility to Plasmodium infection," *Molecular Ecology*, vol. 25, pp. 1494–1510, apr 2016.

[134] E. Erlank, L. L. Koekemoer, and M. Coetzee, "The importance of morphological identification of African anopheline mosquitoes (Diptera: Culicidae) for malaria control programmes," *Malaria Journal*, vol. 17, p. 43, jan 2018.

[135] A. Wiebe, J. Longbottom, K. Gleave, F. M. Shearer, M. E. Sinka, N. C. Massey, E. Cameron, S. Bhatt, P. W. Gething, J. Hemingway, *et al.*, "Geographical distributions of african malaria vector sibling species and evidence for insecticide resistance," *Malaria journal*, vol. 16, no. 1, pp. 1–10, 2017.

[136] M. Coetzee, R. H. Hunt, R. Wilkerson, A. Della Torre, M. B. Coulibaly, N. J. Besansky, A. D. Torre, M. B. Coulibaly, and N. J. Besansky, "Anopheles coluzzii and anopheles amharicus, new members of the anopheles gambiae complex," *Zootaxa*, vol. 3619, no. 3, pp. 246–274, 2013.

[137] M. Coetzee, "Distribution of the african malaria vectors of the anopheles gambiae complex.," *The American journal of tropical medicine and hygiene*, vol. 70, no. 2, pp. 103–104, 2004.

[138] E. O. Ogola, U. Fillinger, I. M. Ondiba, J. Villinger, D. K. Masiga, B. Torto, and D. P. Tchouassi, "Insights into malaria transmission among Anopheles funestus mosquitoes, Kenya," *Parasites and Vectors*, vol. 11, pp. 1–10, nov 2018.

[139] A. Burke, L. Dandalo, G. Munhenga, Y. Dahan-Moss, F. Mbokazi, S. Ngxongo, M. Coetzee, L. Koekemoer, and B. Brooke, "A new malaria vector mosquito in south africa," *Scientific reports*, vol. 7, no. 1, pp. 1–5, 2017.

[140] B. S. Laurent, M. Cooke, S. M. Krishnankutty, P. Asih, J. D. Mueller, S. Kahindi, E. Ayoma, R. M. Oriango, J. Thumloup, C. Drakeley, *et al.*, "Molecular characterization reveals diverse and unknown malaria vectors in the western kenyan highlands," *The American journal of tropical medicine and hygiene*, vol. 94, no. 2, p. 327, 2016.

[141] J. C. Stevenson and D. E. Norris, "Implicating cryptic and novel anophelines as malaria vectors in africa," *Insects*, vol. 8, no. 1, p. 1, 2017.

[142] B. J. White, F. H. Collins, and N. J. Besansky, "Evolution of anopheles gambiae in relation to humans and malaria," *Annual review of ecology, evolution, and systematics*, vol. 42, pp. 111–132, 2011.

[143] P. E. Cook, C. J. Mcmeniman, and S. L. O. Neill, "Modifying Insect Population Age Structure," *Transgenesis and the Management of Vector-Borne Disease*, pp. 126–140, 2008.

[144] L. D. Kramer and A. T. Ciota, "Dissecting vectorial capacity for mosquito-borne viruses," *Current Opinion in Virology*, vol. 15, pp. 112–118, dec 2015.

[145] A. Chan, L. P. Chiang, H. C. Hapuarachchi, C. H. Tan, S. C. Pang, R. Lee, K. S. Lee, L. C. Ng, and S. G. Lam-Phua, "DNA barcoding: Complementing morphological identification of mosquito species in Singapore," *Parasites and Vectors*, vol. 7, no. 1, 2014.

[146] M. A. M. Sallum, R. G. Obando, N. Carrejo, and R. C. Wilkerson, "Identification keys to the Anopheles mosquitoes of South America (Diptera: Culicidae). IV. Adult females," *Parasites and Vectors*, vol. 13, pp. 1–14, nov 2020.

[147] J. A. Scott, W. G. Brogdon, and F. H. Collins, "Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction," *American Journal of Tropical Medicine and Hygiene*, vol. 49, pp. 520–529, oct 1993.

[148] C. Bass, M. S. Williamson, C. S. Wilding, M. J. Donnelly, and L. M. Field, "Identification of the main malaria vectors in the Anopheles gambiae species complex using a TaqMan real-time PCR assay," *Malaria Journal*, vol. 6, p. 155, dec 2007.

[149] J. N. Fernandes, L. M. B. Dos Santos, T. Chouin-Carneiro, M. G. Pavan, G. A. Garcia, M. R. David, J. C. Beier, F. E. Dowell, R. Maciel-De-Freitas, and M. T. Sikulu-Lord, "Rapid, noninvasive detection of Zika virus in Aedes aegypti mosquitoes by near-infrared spectroscopy," *Science Advances*, vol. 4, may 2018.

[150] S. B. Vezenegho, C. Bass, M. Puinean, M. S. Williamson, L. M. Field, M. Coetzee, and L. L. Koekemoer, "Development of multiplex real-time PCR assays for identification of members of the Anopheles funestus species group," *Malaria Journal*, vol. 8, pp. 1–9, dec 2009.

[151] J. B. Silver, *Mosquito ecology: field sampling methods*. springer science & business media, 2007.

[152] R. Perry, "Malaria in the jeypore hill tract and adjoining coastland," *Paludism*, vol. 5, p. 32, 1912.

[153] P. S. Corbet, "Recognition of nulliparous mosquitoes without dissection," *Nature*, vol. 187, no. 4736, pp. 525–526, 1960.

[154] B. Kay, "Age structure of populations of culex annulirostris (diptera: Culicidae) at kowanyama and charleville, queensland," *Journal of Medical Entomology*, vol. 16, no. 4, pp. 309–316, 1979.

[155] A. Suzuki, Y. Tsuda, M. Takagi, and Y. Wada, "Seasonal observation on some population attributes of aedes albopictus," *Tropical medicine*, vol. 35, no. 3, pp. 91–99, 1993.

[156] J. D. Gillett, "Age analysis in the biting-cycle of the mosquito taeniorhynchus (mansonioides) africanus theobald, based on the presence of parasitic mites," *Annals of Tropical Medicine and Parasitology*, vol. 51, no. 2, pp. 151–158, 1957.

[157] P. S. Corbet, "The reliability of parasitic water-mites (Hydracarina) as indicators of physiological age in mosquitoes (Diptera: Culicidae)," *Entomologia Experimentalis et Applicata*, vol. 6, no. 3, pp. 215–233, 1963.

[158] A. McCrae, "The association between larval parasitic water mites (hydracarina) and anopheles implexus (theobald)(diptera, culicidae)," *Bulletin of Entomological Research*, vol. 66, no. 4, pp. 633–650, 1976.

[159] S. Biswas, B. L. Wattal, D. Tyagi, and K. Kumar, "Limitation of larval parasitic water mite infestation in age-gradation of adult Anopheles.," *Journal of Communicable Diseases*, vol. 12, no. 4, pp. 214–215, 1980.

[160] L. E. Hugo, S. Quick-miles, B. H. Kay, and P. A. Ryan, "Evaluations of Mosquito Age Grading Techniques Based on Morphological Changes," *Journal of Medical Entomology*, vol. 45, no. 3, pp. 353–369, 2008.

[161] W. S. Romoser, R. M. Moll, A. C. Moncayo, and K. Lerdthusnee, "The occurrence and fate of the meconium and meconial peritrophic membranes in pupal and adult mosquitoes (Diptera: Culicidae)," *Journal of Medical Entomology*, vol. 37, pp. 893–896, nov 2000.

[162] J. G. Hitchcock Jr, "Age composition of a natural population of anopheles quadrimaculatus say (diptera: Culicidae) in maryland, usa," *Journal of Medical Entomology*, vol. 5, no. 1, pp. 125–134, 1968.

[163] Y. Schlein and N. G. Gratz, "Determination of the age of some anopheline mosquitos by daily growth layers of skeletal apodemes," *Bulletin of the World Health Organization*, vol. 49, no. 4, pp. 371–375, 1973.

[164] T. S. Detinova, D. S. Bertram, and W. H. Organization, "Age-grouping methods in diptera of medical importance, with special reference to some vectors of malaria / T. S. Detinova ; [with] an Annex on the ovary and ovarioles of mosquitos (with glossary) by D. S. Bertram," 1962.

[165] M. T. Gillies and T. J. Wilkes, "A study of the age-composition of populations of anopheles gambiae giles and a. funestus giles in north-eastern tanzania," *Bulletin of entomological research*, vol. 56, no. 2, pp. 237–262, 1965.

[166] V. Polovodova, "The determination of the physiological age of female," 1949.

[167] M. Service, *Medical entomology for students, fourth edition.* 2008.

[168] W. N. Beklemishev, T. S. Detinova, and V. P. Polovodova, "Determination of physiological age in anophelines and of age distribution in anopheline populations in the USSR.," *Bulletin of the World Health Organization*, vol. 21, pp. 223–232, 1959.

[169] A. Clements and G. A. Kerkut, "The Physiology of Mosquitoes: International Series of Monographs," *PERGAMON PRESS LTD.*, vol. 17, p. 410, 1963.

[170] F. J. Lardeux, R. H. Tejerina, V. Quispe, and T. K. Chavez, "A physiological time analysis of the duration of the gonotrophic cycle of Anopheles pseudopunctipennis and its implications for malaria transmission in Bolivia," *Malaria Journal*, vol. 7, pp. 1–17, jul 2008.

[171] G. L. Rúa, M. L. Quiñones, I. D. Vélez, J. S. Zuluaga, W. Rojas, G. Poveda, and D. Ruiz, "Laboratory estimation of the effects of increasing temperatures on the duration of gonotrophic cycle of Anopheles albimanus (Diptera: Culicidae)," *Memorias do Instituto Oswaldo Cruz*, vol. 100, pp. 515–520, aug 2005.

[172] T. P. Agyekum, P. K. Botwe, J. Arko-Mensah, I. Issah, A. A. Acquah, J. N. Hogarh, D. Dwomoh, T. Robins, and J. N. Fobil, "A systematic review of the effects of temperature on anopheles mosquito development and survival: Implications for malaria control in a future warmer climate," 2021.

[173] T. Q. HOC and J. D. CHARLWOOD, "Age determination of Aedes cantans using the ovarian oil injection technique," *Medical and Veterinary Entomology*, vol. 4, no. 2, pp. 227–233, 1990.

[174] C. C. Draper and G. Davidson, "A new method of estimating the survival-rate of anopheline mosquitoes in nature [13]," *Nature*, vol. 172, no. 4376, p. 503, 1953.

[175] T. Hoc and G. A. Schaub, "Ovariolar 'basal body' development and physiological age of the mosquito Aedes aegypti," *Medical and Veterinary Entomology*, vol. 9, pp. 9–15, jan 1995.

[176] N. Liu, "Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and Research Directions," *Annual Review of Entomology*, vol. 60, pp. 537–559, jan 2015.

[177] J. E. Casida and K. A. Durkin, "Neuroactive insecticides: targets, selectivity, resistance, and secondary effects," *Annual review of entomology*, vol. 58, pp. 99–117, 2013.

[178] V. Balabanidou, L. Grigoraki, and J. Vontas, "Insect cuticle: a critical determinant of insecticide resistance," jun 2018.

[179] V. Balabanidou, M. Kefi, M. Aivaliotis, V. Koidou, J. R. Girotti, S. J. Mijailovsky, M. P. Juárez, E. Papadogiorgaki, G. Chalepakis, A. Kampouraki, C. Nikolaou, H. Ranson, and J. Vontas, "Mosquitoes cloak their legs to resist insecticides," *Proceedings of the Royal Society B: Biological Sciences*, vol. 286, no. 1907, p. 20191091, 2019.

[180] C. Bass and C. M. Jones, "Mosquitoes boost body armor to resist insecticide attack," *Proceedings of the National Academy of Sciences*, vol. 113, no. 33, pp. 9145–9147, 2016.

[181] G. A. Yahouédo, F. Chandre, M. Rossignol, C. Ginibre, V. Balabanidou, N. G. A. Mendez, O. Pigeon, J. Vontas, and S. Cornelie, "Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector Anopheles gambiae," *Scientific Reports*, vol. 7, p. 11091, dec 2017.

[182] A. Lynd, A. Oruni, A. E. Van'T Hof, J. C. Morgan, L. B. Naego, D. Pipini, K. A. O'Kines, T. L. Bobanga, M. J. Donnelly, and D. Weetman, "Insecticide resistance in Anopheles gambiae from the northern Democratic Republic of Congo, with extreme knockdown resistance (kdr) mutation frequencies revealed by a new diagnostic assay," *Malaria Journal*, vol. 17, pp. 1–8, nov 2018.

[183] J. Yin, F. Yamba, C. Zheng, S. Zhou, S. J. Smith, L. Wang, H. Li, Z. Xia, and N. Xiao, "Molecular Detection of Insecticide Resistance Mutations in Anopheles gambiae from Sierra Leone Using Multiplex SNaPshot and Sequencing," *Frontiers in Cellular and Infection Microbiology*, vol. 11, p. 778, aug 2021.

[184] S. S. Ibrahim, J. M. Riveron, J. Bibby, H. Irving, C. Yunta, M. J. Paine, and C. S. Wondji, "Allelic variation of cytochrome p450s drives resistance to bednet insecticides in a major malaria vector," *PLoS genetics*, vol. 11, no. 10, p. e1005618, 2015.

[185] S. N. Mitchell, D. J. Rigden, A. J. Dowd, F. Lu, C. S. Wilding, D. Weetman, S. Dadzie, A. M. Jenkins, K. Regna, P. Boko, L. Djogbenou, M. A. T. Muskavitch, H. Ranson, M. J. I. Paine, O. Mayans, and M. J. Donnelly, "Metabolic and Target-Site Mechanisms Combine to Confer Strong DDT Resistance in Anopheles gambiae," *PLOS ONE*, vol. 9, no. 3, pp. 1–10, 2014.

[186] J. M. Riveron, C. Yunta, S. S. Ibrahim, R. Djouaka, H. Irving, B. D. Menze, H. M. Ismail, J. Hemingway, H. Ranson, A. Albert, *et al.*, "A single mutation in the gste2 gene allows tracking of metabolically based insecticide resistance in a major malaria vector," *Genome biology*, vol. 15, no. 2, pp. 1–20, 2014.

[187] M. J. Donnelly, A. T. Isaacs, and D. Weetman, "Identification, Validation, and Application of Molecular Diagnostics for Insecticide Resistance in Malaria Vectors," *Trends in Parasitology*, vol. 32, pp. 197–206, mar 2016.

[188] World Health Organization (WHO), *Test Procedures for Insecticide Resistance Monitoring in Malaria Vector Mosquitoes*. 2013.

[189] I. Potamitis and I. Rigakis, "Measuring the fundamental frequency and the harmonic properties of the wingbeat of a large number of mosquitoes in flight using 2D optoacoustic sensors," *Applied Acoustics*, vol. 109, pp. 54–60, aug 2016.

[190] T. H. Ouyang, E. C. Yang, J. A. Jiang, and T. T. Lin, "Mosquito vector monitoring system based on optical wingbeat classification," *Computers and Electronics in Agriculture*, vol. 118, pp. 47–55, 10 2015.

[191] A. P. Genoud, R. Basistyy, G. M. Williams, and B. P. Thomas, "Optical remote sensing for monitoring flying mosquitoes, gender identification and discussion on species identification," *Applied Physics B: Lasers and Optics*, vol. 124, p. 46, mar 2018.

[192] H. Mukundarajan, F. J. H. Hol, E. A. Castillo, C. Newby, and M. Prakash, "Using mobile phones as acoustic sensors for high-throughput mosquito surveillance," *eLife*, vol. 6, oct 2017.

[193] A. Goodwin, S. Padmanabhan, S. Hira, M. Glancey, M. Slinowsky, R. Immidisetti, L. Scavo, J. Brey, B. M. M. S. Sudhakar, T. Ford, *et al.*, "Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection," *Scientific reports*, vol. 11, no. 1, pp. 1–15, 2021.

[194] D. Motta, A. Á. B. Santos, I. Winkler, B. A. S. Machado, D. A. D. I. Pereira, A. M. Cavalcanti, E. O. L. Fonseca, F. Kirchner, and R. Badaró, "Application of convolutional neural networks for classification of adult mosquitoes in the field," *PLoS ONE*, vol. 14, p. e0210829, jan 2019.

[195] J. Park, D. I. Kim, B. Choi, W. Kang, and H. W. Kwon, "Classification and morphological analysis of vector mosquitoes using deep convolutional neural networks," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[196] V. S. Mayagaya, K. Michel, M. Q. Benedict, G. F. Killeen, R. A. Wirtz, H. M. Ferguson, and F. E. Dowell, "Non-destructive determination of age and species of Anopheles gambiae s.l. using near-infrared spectroscopy," *American Journal of Tropical Medicine and Hygiene*, vol. 81, no. 4, pp. 622–630, 2009.

[197] M. Sikulu, G. F. Killeen, L. E. Hugo, P. A. Ryan, K. M. Dowell, R. A. Wirtz, S. J. Moore, and F. E. Dowell, "Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors," *Parasites & Vectors*, vol. 3, no. 1, p. 49, 2010.

[198] B. Lambert, M. T. Sikulu-Lord, V. S. Mayagaya, G. Devine, F. Dowell, and T. S. Churcher, "Monitoring the Age of Mosquito Populations Using Near-Infrared Spectroscopy," *Scientific Reports*, vol. 8, p. 5274, dec 2018.

[199] D. J. Siria, R. Sanou, J. Mitton, E. P. Mwanga, A. Niang, I. Sare, P. C. Johnson, G. M. Foster, A. M. Belem, K. Wynne, R. Murray-Smith, H. M. Ferguson, M. González-Jiménez, S. A. Babayan, A. Diabaté, F. O. Okumu, and F. Baldini, "Rapid age-grading and species identification of natural mosquitoes for malaria surveillance," *Nature Communications*, vol. 13, pp. 1–9, mar 2022.

[200] M. V. Braga, Z. T. Pinto, M. M. de Carvalho Queiroz, N. Matsumoto, and G. J. Blomquist, "Cuticular hydrocarbons as a tool for the identification of insect species: Puparial cases from Sarcophagidae," *Acta Tropica*, vol. 128, no. 3, pp. 479–485, 2013.

[201] B. Caputo, F. R. Dani, G. L. Horne, V. Petrarca, S. Turillazzi, M. Coluzzi, A. A. Priestman, and A. Della Torre, "Identification and composition of cuticular hydrocarbons of the major Afrotropical malaria vector Anopheles gambiae s.s. (Diptera: Culicidae): Analysis of sexual dimorphism and age-related changes," *Journal of Mass Spectrometry*, vol. 40, no. 12, pp. 1595–1604, 2005.

[202] K. H. Lockey, "Insect cuticular lipids," *Comparative Biochemistry and Physiology – Part B: Biochemistry and*, vol. 81, no. 2, pp. 263–273, 1985.

[203] C. S. Chen, M. S. Mulla, R. B. March, and J. D. Chaney, "Cuticular hydrocarbon patterns in Culex quinquefasciatus as influenced by age, sex, and geography.," *Journal of Vector Ecology*, vol. 22, no. 1, pp. 1–98, 1997.

[204] M. L. Desena, J. M. Clark, J. D. Edman, S. B. Symington, T. W. Scott, G. G. Clark, and T. M. Peters, "Potential for aging female Aedes aegypti (Diptera: Culicidae) by gas chromatographic analysis of cuticular hydrocarbons, including a field evaluation," *Journal of Medical Entomology*, vol. 36, no. 6, pp. 811–823, 1999.

[205] B. B. Gerade, S. H. Lee, T. W. Scott, J. D. Edman, L. C. Harrington, S. Kitthawee, J. W. Jones, and J. M. Clark, "Field validation of Aedes aegypti (Diptera: Culicidae) age estimation by analysis of cuticular hydrocarbons.," *Journal of Medical Entomology*, vol. 41, no. 2, pp. 231–238, 2004.

[206] M. H. Wang, O. Marinotti, A. A. James, E. Walker, J. Githure, and G. Yan, "Genome-wide patterns of gene expression during aging in the African malaria vector Anopheles gambiae," *PLoS ONE*, vol. 5, p. e13359, oct 2010.

[207] M.-H. Wang, O. Marinotti, D. Zhong, A. A. James, E. Walker, T. Guda, E. J. Kweka, J. Githure, and G. Yan, "Gene Expression-Based Biomarkers for Anopheles gambiae Age Grading," *PLoS ONE*, vol. 8, p. e69439, jul 2013.

[208] P. E. Cook, L. E. Hugo, I. Iturbe-Ormaetxe, C. R. Williams, S. F. Chenoweth, S. A. Ritchie, P. A. Ryan, B. H. Kay, M. W. Blows, and S. L. O'Neill, "Predicting the age of

mosquitoes using transcriptional profiles," *Nature Protocols*, vol. 2, no. 11, pp. 2796–2806, 2007.

[209] T. K. Joy, E. H. Jeffrey Gutierrez, K. Ernst, K. R. Walker, Y. Carriere, M. Torabi, and M. A. Riehle, "Aging Field Collected Aedes aegypti to Determine Their Capacity for Dengue Transmission in the Southwestern United States," *PLoS ONE*, vol. 7, p. e46946, oct 2012.

[210] M. T. Sikulu, J. Monkman, K. A. Dave, M. L. Hastie, P. E. Dale, R. L. Kitching, G. F. Killeen, B. H. Kay, J. J. Gorman, and L. E. Hugo, "Proteomic changes occurring in the malaria mosquitoes Anopheles gambiae and Anopheles stephensi during aging," *Journal of Proteomics*, vol. 126, pp. 234–244, aug 2015.

[211] C. Nabet, A. Chaline, J. F. Franetich, J. Y. Brossas, N. Shahmirian, O. Silvie, X. Tannier, and R. Piarroux, "Prediction of malaria transmission drivers in Anopheles mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry," *Scientific Reports*, vol. 10, pp. 1–13, jul 2020.

[212] L. E. Hugo, J. Monkman, K. A. Dave, L. F. Wockner, G. W. Birrell, E. L. Norris, V. J. Kienzle, M. T. Sikulu, P. A. Ryan, J. J. Gorman, and B. H. Kay, "Proteomic Biomarkers for Ageing the Mosquito Aedes aegypti to Determine Risk of Pathogen Transmission," *PLoS ONE*, vol. 8, no. 3, 2013.

[213] M. T. Sikulu, S. Majambere, B. O. Khatib, A. S. Ali, L. E. Hugo, and F. E. Dowell, "Using a near-infrared spectrometer to estimate the age of Anopheles mosquitoes exposed to pyrethroids," *PLoS ONE*, vol. 9, no. 3, 2014.

[214] B. J. Krajacich, J. I. Meyers, H. Alout, R. K. Dabiré, F. E. Dowell, and B. D. Foy, "Analysis of near infrared spectra for age-grading of wild populations of Anopheles gambiae," *Parasites & Vectors*, vol. 10, no. 1, p. 552, 2017.

[215] A. J. Ntamatungiro, V. S. Mayagaya, S. Rieben, S. J. Moore, F. E. Dowell, and M. F. Maia, "The influence of physiological status on age prediction of Anopheles arabiensis using near infra-red spectroscopy," *Parasit Vectors*, vol. 6, no. 1, p. 298, 2013.

[216] F. E. Dowell, A. E. M. Noutcha, and K. Michel, "Short Report: The effect of preservation methods on predicting mosquito age by Near Infrared Spectroscopy," *Am J Trop Med Hyg*, vol. 85, no. 6, pp. 1092–1096, 2011.

[217] O. T. Ong, E. A. Kho, P. M. Esperança, C. Freebairn, F. E. Dowell, G. J. Devine, and T. S. Churcher, "Ability of near-infrared spectroscopy and chemometrics to predict the age of mosquitoes reared under different conditions," *Parasites and Vectors*, vol. 13, pp. 1–10, mar 2020.

[218] M. P. Milali, S. S. Kiware, N. J. Govella, F. Okumu, N. Bansal, S. Bozdag, J. D. Charlwood, M. F. Maia, S. B. Ogoma, F. E. Dowell, G. F. Corliss, M. T. Sikulu-Lord, and R. J.

Povinelli, "An autoencoder and artificial neural network-based method to estimate parity status of wild mosquitoes from near-infrared spectra," *PLoS ONE*, vol. 15, p. e0234557, jun 2020.

[219] M. P. Milali, M. T. Sikulu-Lord, S. S. Kiware, F. E. Dowell, G. F. Corliss, and R. J. Povinelli, "Age grading an. gambiae and an. arabiensis using near infrared spectra and artificial neural networks," *PloS one*, vol. 14, no. 8, p. e0209451, 2019.

[220] M. T. Sikulu-Lord, G. J. Devine, L. E. Hugo, and F. E. Dowell, "First report on the application of near-infrared spectroscopy to predict the age of Aedes albopictus Skuse," *Scientific Reports*, vol. 8, p. 9590, dec 2018.

[221] M. T. Sikulu-Lord, M. F. Maia, M. P. Milali, M. Henry, G. Mkandawile, E. A. Kho, R. A. Wirtz, L. E. Hugo, F. E. Dowell, and G. J. Devine, "Rapid and Non-destructive Detection and Identification of Two Strains of Wolbachia in Aedes aegypti by Near-Infrared Spectroscopy," *PLoS Neglected Tropical Diseases*, vol. 10, no. 6, pp. 1–12, 2016.

[222] A. Khoshmanesh, D. Christensen, D. Perez-Guaita, I. Iturbe-Ormaetxe, S. L. O'Neill, D. McNaughton, and B. R. Wood, "Screening of Wolbachia Endosymbiont Infection in Aedes aegypti Mosquitoes Using Attenuated Total Reflection Mid-Infrared Spectroscopy," *Analytical Chemistry*, vol. 89, no. 10, pp. 5285–5293, 2017.

[223] D. Wang, J. Yang, J. Pandya, J. M. Clark, L. C. Harrington, C. C. Murdock, and L. He, "Quantitative age grading of mosquitoes using surface-enhanced Raman spectroscopy," *Analytical Science Advances*, vol. 3, pp. 47–53, feb 2022.

[224] L. Sroute, B. D. Byrd, and S. W. Huffman, "Classification of Mosquitoes with Infrared Spectroscopy and Partial Least Squares-Discriminant Analysis," *Applied Spectroscopy*, vol. 74, no. 8, pp. 900–912, 2020.

[225] M. T. Sikulu-Lord, M. P. Milali, M. Henry, R. A. Wirtz, L. E. Hugo, F. E. Dowell, and G. J. Devine, "Near-Infrared Spectroscopy, a Rapid Method for Predicting the Age of Male and Female Wild-Type and Wolbachia Infected Aedes aegypti," *PLoS Neglected Tropical Diseases*, vol. 10, no. 10, pp. 1–11, 2016.

[226] L. M. Santos, M. Mutsaers, G. A. Garcia, M. R. David, M. G. Pavan, M. T. Petersen, J. Corrêa-Antônio, D. Couto-Lima, L. Maes, F. Dowell, A. Lord, M. Sikulu-Lord, and R. Maciel-de Freitas, "High throughput estimates of Wolbachia, Zika and chikungunya infection in Aedes aegypti by near-infrared spectroscopy to improve arbovirus surveillance," *Communications Biology*, vol. 4, pp. 1–9, jan 2021.

[227] A. Tátila-Ferreira, G. A. Garcia, L. M. dos Santos, M. G. Pavan, C. J. Carlos, J. C. Victoriano, R. da Silva-Junior, J. R. dos Santos-Mallet, T. Verly, C. Britto, M. T. Sikulu-Lord, and R. Maciel-de Freitas, "Near infrared spectroscopy accurately detects Trypanosoma cruzi

non-destructively in midguts, rectum and excreta samples of Triatoma infestans," *Scientific Reports*, vol. 11, dec 2021.

[228] S. Roy, D. Perez-Guaita, D. W. Andrew, J. S. Richards, D. McNaughton, P. Heraud, and B. R. Wood, "Simultaneous ATR-FTIR Based Determination of Malaria Parasitemia, Glucose and Urea in Whole Blood Dried onto a Glass Slide," *Analytical Chemistry*, vol. 89, no. 10, pp. 5238–5245, 2017.

[229] E. P. Mwanga, E. G. Minja, E. Mrimi, M. G. Jiménez, J. K. Swai, S. Abbasi, H. S. Ngowo, D. J. Siria, S. Mapua, C. Stica, M. F. Maia, A. Olotu, M. T. Sikulu-Lord, F. Baldini, H. M. Ferguson, K. Wynne, P. Selvaraj, S. A. Babayan, and F. O. Okumu, "Detection of malaria parasites in dried human blood spots using mid-infrared spectroscopy and logistic regression analysis," *Malaria Journal*, vol. 18, p. 341, oct 2019.

[230] E. P. Mwanga, S. A. Mapua, D. J. Siria, H. S. Ngowo, F. Nangacha, J. Mgando, F. Baldini, M. González Jiménez, H. M. Ferguson, K. Wynne, P. Selvaraj, S. A. Babayan, and F. O. Okumu, "Using mid-infrared spectroscopy and supervised machine-learning to identify vertebrate blood meals in the malaria vector, Anopheles arabiensis," *Malaria Journal*, vol. 18, p. 187, dec 2019.

[231] M. P. Milali, M. T. Sikulu-Lord, S. S. Kiware, F. E. Dowell, G. F. Corliss, and R. J. Povinelli, "Age grading An. gambiae and An. arabiensis using near infrared spectra and artificial neural networks," *PLoS ONE*, vol. 14, p. e0209451, aug 2019.

[232] W. H. Organization *et al.*, "World malaria report 2021," 2021.

[233] G. Benelli and J. C. Beier, "Current vector control challenges in the fight against malaria," oct 2017.

[234] J. A. Ogunah, J. O. Lalah, and K. W. Schramm, "Malaria vector control strategies. What is appropriate towards sustainable global eradication?," dec 2020.

[235] D. E. Impoinvil, S. Ahmad, A. Troyo, J. Keating, A. K. Githeko, C. M. Mbogo, L. Kibe, J. I. Githure, A. M. Gad, A. N. Hassan, L. Orshan, A. Warburg, O. Calderón-Arguedas, V. M. Sánchez-Loría, R. Velit-Suarez, D. D. Chadee, R. J. Novak, and J. C. Beier, "Comparison of mosquito control programs in seven urban sites in Africa, the Middle East, and the Americas," *Health Policy*, vol. 83, no. 2-3, pp. 196–212, 2007.

[236] T. L. Russell, R. Farlow, M. Min, E. Espino, A. Mnzava, and T. R. Burkot, "Capacity of National Malaria Control Programmes to implement vector surveillance: a global analysis," *Malaria Journal*, vol. 19, pp. 1–9, nov 2020.

[237] World Health Organization, "Dengue and severe dengue - Fact Sheet," *April*, pp. 1–5, 2017.

[238] S. Camara, A. A. Koffi, L. P. Ahoua Alou, K. Koffi, J. P. K. Kabran, A. Koné, M. F. Koffi, R. N'Guessan, and C. Pennetier, "Mapping insecticide resistance in Anopheles gambiae (s.l.) from Côte d'Ivoire," *Parasites and Vectors*, vol. 11, pp. 1–11, jan 2018.

[239] C. Stica, C. L. Jeffries, S. R. Irish, Y. Barry, D. Camara, I. Yansane, M. Kristan, T. Walker, and L. A. Messenger, "Characterizing the molecular and metabolic mechanisms of insecticide resistance in Anopheles gambiae in Faranah, Guinea," *Malaria Journal*, vol. 18, pp. 1–15, jul 2019.

[240] T. S. Churcher, N. Lissenden, J. T. Griffin, E. Worrall, and H. Ranson, "The impact of pyrethroid resistance on the efficacy and effectiveness of bednets for malaria control in Africa," *eLife*, vol. 5, aug 2016.

[241] S. Sougoufara, M. Harry, S. Doucouré, P. M. Sembène, and C. Sokhna, "Shift in species composition in the Anopheles gambiae complex after implementation of long-lasting insecticidal nets in Dielmo, Senegal," *Medical and veterinary entomology*, vol. 30, pp. 365–368, sep 2016.

[242] H. D. Mawejje, M. Kilama, S. P. Kigozi, A. K. Musiime, M. Kamya, J. Lines, S. W. Lindsay, D. Smith, G. Dorsey, M. J. Donnelly, and S. G. Staedke, "Impact of seasonality and malaria control interventions on Anopheles density and species composition from three areas of Uganda with differing malaria endemicity," *Malaria Journal*, vol. 20, pp. 1–13, mar 2021.

[243] J. I. Meyers, S. Pathikonda, Z. R. Popkin-Hall, M. C. Medeiros, G. Fuseini, A. Matias, G. Garcia, H. J. Overgaard, V. Kulkarni, V. P. Reddy, C. Schwabe, J. Lines, I. Kleinschmidt, and M. A. Slotman, "Increasing outdoor host-seeking in Anopheles gambiae over 6 years of vector control on Bioko Island," *Malaria Journal*, vol. 15, pp. 1–13, apr 2016.

[244] A. K. Musiime, D. L. Smith, M. Kilama, J. Rek, E. Arinaitwe, J. I. Nankabirwa, M. R. Kamya, M. D. Conrad, G. Dorsey, A. M. Akol, S. G. Staedke, S. W. Lindsay, and J. P. Egonyu, "Impact of vector control interventions on malaria transmission intensity, outdoor vector biting rates and Anopheles mosquito species composition in Tororo, Uganda," *Malaria Journal*, vol. 18, pp. 1–9, dec 2019.

[245] G. F. Killeen, J. M. Marshall, S. S. Kiware, A. B. South, L. S. Tusting, P. P. Chaki, and N. J. Govella, "Measuring, manipulating and exploiting behaviours of adult mosquitoes to optimise malaria vector control impact," *BMJ Global Health*, vol. 2, no. 2, p. 212, 2017.

[246] S. Sougoufara, S. Doucouré, P. M. Sembéne, M. Harry, and C. Sokhna, "Challenges for malaria vector control in sub-Saharan Africa: Resistance and behavioral adaptations in Anopheles populations," 2017.

[247] S. Sougoufara, E. C. Ottih, and F. Tripet, "The need for new vector control approaches targeting outdoor biting Anopheline malaria vector communities," jun 2020.

[248] T. Degefa, D. Yewhalaw, G. Zhou, M. C. Lee, H. Atieli, A. K. Githeko, and G. Yan, "Indoor and outdoor malaria vector surveillance in western Kenya: Implications for better understanding of residual transmission," *Malaria Journal*, vol. 16, pp. 1–13, nov 2017.

[249] P. G. Pinda, C. Eichenberger, H. S. Ngowo, D. S. Msaky, S. Abbasi, J. Kihonda, H. Bwanaly, and F. O. Okumu, "Comparative assessment of insecticide resistance phenotypes in two major malaria vectors, Anopheles funestus and Anopheles arabiensis in south-eastern Tanzania," *Malaria Journal*, vol. 19, pp. 1–11, nov 2020.

[250] M. Gillies and M. Coetzee, "A supplement to the anophelinae of africa south of the sahara," *Publ S Afr Inst Med Res*, vol. 55, pp. 1–143, 1987.

[251] B. De Meillon *et al.*, "Illustrated keys to the full-grown larvae and adults of south african anopheline mosquitos.," *Illustrated Keys to the full-grown Larvae and Adults of South African Anopheline Mosquitos.*, no. 28, 1931.

[252] M. Coetzee, "Key to the females of Afrotropical Anopheles mosquitoes (Diptera: Culicidae)," *Malaria Journal*, vol. 19, pp. 1–20, feb 2020.

[253] J. Chabi, A. Van't Hof, L. K. N'dri, A. Datsomor, D. Okyere, H. Njoroge, D. Pipini, M. P. Hadi, D. K. De Souza, T. Suzuki, S. K. Dadzie, and H. P. Jamet, "Rapid high throughput SYBR green assay for identifying the malaria vectors Anopheles arabiensis, Anopheles coluzzii and Anopheles gambiae s.s. Giles," *PLoS ONE*, vol. 14, p. e0215669, apr 2019.

[254] M. Bonizzoni, Y. Afrane, and G. Yan, "Loop-mediated isothermal amplification (LAMP) for rapid identification of Anopheles gambiae and Anopheles arabiensis mosquitoes," *American Journal of Tropical Medicine and Hygiene*, vol. 81, pp. 1030–1034, dec 2009.

[255] G. C. Muller, A. Junnila, M. M. Traore, S. F. Traore, S. Doumbia, F. Sissoko, S. M. Dembele, Y. Schlein, K. L. Arheart, E. E. Revay, V. D. Kravchenko, A. Witt, and J. C. Beier, "The invasive shrub Prosopis juliflora enhances the malaria parasite transmission capacity of Anopheles mosquitoes: A habitat manipulation experiment," *Malaria Journal*, vol. 16, pp. 1–9, jul 2017.

[256] B. Emidi, W. N. Kisinza, and F. W. Mosha, "Impact of non-pyrethroid insecticide treated durable wall lining on age structure of malaria vectors in Muheza, Tanzania," *BMC Research Notes*, vol. 10, pp. 1–5, dec 2017.

[257] J. C. Beier, "Malaria parasite development in mosquitoes," *Annual review of entomology*, vol. 43, no. 1, pp. 519–543, 1998.

[258] W. H. Organization *et al.*, *Indoor residual spraying: an operational manual for indoor residual spraying (IRS) for malaria transmission control and elimination*. World Health Organization, 2015.

[259] D. L. Smith, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie, "Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens," *PLoS Pathogens*, vol. 8, p. e1002588, apr 2012.

[260] M. T. White, J. T. Griffin, T. S. Churcher, N. M. Ferguson, M. G. Basáñez, and A. C. Ghani, "Modelling the impact of vector control interventions on Anopheles gambiae population dynamics," *Parasites and Vectors*, vol. 4, pp. 1–14, jul 2011.

[261] A. Hughes, N. Lissenden, M. Viana, K. H. Toé, and H. Ranson, "Anopheles gambiae populations from Burkina Faso show minimal delayed mortality after exposure to insecticide-treated nets," *Parasites and Vectors*, vol. 13, pp. 1–11, jan 2020.

[262] D. L. Smith and F. E. McKenzie, "Statics and dynamics of malaria infection in Anopheles mosquitoes," jun 2004.

[263] E. Suarez, H. P. Nguyen, I. P. Ortiz, K. J. Lee, S. B. Kim, J. Krzywinski, and K. A. Schug, "Matrix-assisted laser desorption/ionization-mass spectrometry of cuticular lipid profiles can differentiate sex, age, and mating status of Anopheles gambiae mosquitoes," *Analytica Chimica Acta*, vol. 706, pp. 157–163, nov 2011.

[264] P. E. Cook, L. E. Hugo, I. Iturbe-Ormaetxe, C. R. Williams, S. F. Chenoweth, S. A. Ritchie, P. A. Ryan, B. H. Kay, M. W. Blows, and S. L. O'Neill, "The use of transcriptional profiles to predict adult mosquito age under field conditions.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 48, pp. 18060–18065, 2006.

[265] V. Balabanidou, A. Kampouraki, M. MacLean, G. J. Blomquist, C. Tittiger, M. P. Juárez, S. J. Mijailovsky, G. Chalepakis, A. Anthousi, A. Lynd, S. Antoine, J. Hemingway, H. Ranson, G. J. Lycett, and J. Vontas, "Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in Anopheles gambiae," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 9268–9273, aug 2016.

[266] K. Mavridis, N. Wipf, S. Medves, I. Erquiaga, P. Müller, and J. Vontas, "Rapid multiplex gene expression assays for monitoring metabolic resistance in the major malaria vector Anopheles gambiae 06 Biological Sciences 0604 Genetics," *Parasites and Vectors*, vol. 12, p. 9, dec 2019.

[267] S. Türker-Kaya and C. W. Huck, "A review of mid-infrared and near-infrared imaging: Principles, concepts and applications in plant tissue analysis," jan 2017.

[268] M. Manley, "Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials," dec 2014.

[269] C. Pasquini, "Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications," 2003.

[270] J. A. Adegoke, K. Kochan, P. Heraud, and B. R. Wood, "A Near-Infrared "Matchbox Size" Spectrometer to Detect and Quantify Malaria Parasitemia," *Analytical Chemistry*, vol. 93, pp. 5451–5458, apr 2021.

[271] P. Bassan, A. Sachdeva, J. Lee, and P. Gardner, "Substrate contributions in micro-ATR of thin samples: Implications for analysis of cells, tissue and biological fluids," *Analyst*, vol. 138, pp. 4139–4146, jun 2013.

[272] S. Depickère, A. G. Ravelo-García, and F. Lardeux, "Chagas disease vectors identification using visible and near-infrared spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 197, p. 103914, feb 2020.

[273] W. K. Reeves, K. H. Peiris, E. J. Scholte, R. A. Wirtz, and F. E. Dowell, "Age-grading the biting midge Culicoides sonorensis using near-infrared spectroscopy," *Medical and Veterinary Entomology*, vol. 24, no. 1, pp. 32–37, 2010.

[274] J. Paliwal, W. Wang, S. Symons, and C. Karunakaran, "Insect species and infestation level determination in stored wheat using near-infrared spectroscopy," *Canadian Biosystems Engineering*, vol. 46, no. 7, pp. 17–24, 2004.

[275] P. Cheewapramong and R. Wehling, "A simplified near–infrared method for detecting internal insect infestation in wheat kernels," in *AACC Abstract Paper*, vol. 368, 2001.

[276] K. K. Kalsa, B. Subramanyam, G. Demissie, A. F. Worku, and N. G. Habtu, "Major insect pests and their associated losses in quantity and quality of farm-stored wheat seed," *Ethiopian Journal of Agricultural Sciences*, vol. 29, no. 2, pp. 71–82, 2019.

[277] S. Fischnaller, F. E. Dowell, A. Lusser, B. C. Schlick-Steiner, and F. M. Steiner, "Non-destructive species identification of Drosophila obscura and D. subobscura (Diptera) using near-infrared spectroscopy," *Fly*, vol. 6, pp. 284–289, oct 2012.

[278] M. F. Maia, M. Kapulu, M. Muthui, M. G. Wagah, H. M. Ferguson, F. E. Dowell, F. Baldini, and L.-R. R. Cartwright, "Detection of Plasmodium falciparum infected Anopheles gambiae using near-infrared spectroscopy," *Malaria Journal*, vol. 18, p. 85, dec 2019.

[279] D. Cozzolino, "An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals," *Food Research International*, vol. 60, pp. 262–265, jun 2014.

[280] M. V. Padalkar and N. Pleshko, "Wavelength-dependent penetration depth of near infrared radiation into cartilage," *Analyst*, vol. 140, pp. 2093–2100, apr 2015.

[281] J. B. Johnson, K. Walsh, and M. Naiker, "Application of infrared spectroscopy for the prediction of nutritional content and quality assessment of faba bean (Vicia faba L.)," *Legume Science*, vol. 2, p. e40, apr 2020.

[282] F. M. Mirabella, *Modern techniques in applied molecular spectroscopy*. Wiley, 1998.

[283] W. T. Wihlborg, J. A. Reffner, S. W. Strand, and F. M. Wasacz, "Reflection Spectroscopy With The FT-IR Microscope," in *7th Intl Conf on Fourier Transform Spectroscopy*, vol. 1145, p. 305, SPIE, dec 1989.

[284] C. Harris, L. Lambrechts, F. Rousset, L. Abate, S. E. Nsango, D. Fontenille, I. Morlais, and A. Cohuet, "Polymorphisms in Anopheles gambiae immune genes associated with natural resistance to plasmodium falciparum," *PLoS Pathogens*, vol. 6, p. 1001112, sep 2010.

[285] J. Williams, L. Flood, G. Praulins, V. A. Ingham, J. Morgan, R. S. Lees, and H. Ranson, "Characterisation of Anopheles strains used for laboratory screening of new vector control products," *Parasites and Vectors*, vol. 12, nov 2019.

[286] M. Gonzalez-Jimenez, S. A. Babayan, P. Khazaeli, M. Doyle, F. Walton, E. Reedy, T. Glew, M. Viana, L. Ranford-Cartwright, A. Niang, D. J. Siria, F. O. Okumu, A. Diabate, H. M. Ferguson, F. Baldini, K. Wynne, M. González-Jiménez, S. A. Babayan, P. Khazaeli, M. Doyle, F. Wal-Ton, E. Reedy, T. Glew, M. Viana, L. Ranford-Cartwright, A. Niang, D. J. Siria, F. O. Okumu, A. Diabaté, H. M. Ferguson, F. Baldini, and K. Wynne, "Prediction of malaria mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning," *bioRxiv*, p. 414342, 2018.

[287] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[288] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, p. e0224365, nov 2019.

[289] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[290] J. J. Jansen, J. Engel, J. Gerretzen, L. Blanchet, E. Szymańska, G. Downey, and L. M. Buydens, "Breaking with trends in pre-processing?," *TrAC Trends in Analytical Chemistry*, vol. 50, pp. 96–106, 2013.

[291] J. Torniainen, I. O. Afara, M. Prakash, J. K. Sarin, L. Stenroth, and J. Töyräs, "Open-source python module for automated preprocessing of near infrared spectroscopic data," *Analytica Chimica Acta*, vol. 1108, pp. 1–9, apr 2020.

[292] V. Machovič, L. Lapčák, M. Havelcová, L. Borecká, M. Novotná, M. Novotná, I. Javůrková, I. Langrová, S. Hájková, A. Brožová, and D. Titěra, "Analysis of European

Honeybee (Apis Mellifera) Wings Using ATR-FTIR and Raman Spectroscopy: A Pilot Study," *Scientia Agriculturae Bohemica*, vol. 48, no. 1, pp. 22–29, 2017.

[293] L. E. Hugo, B. H. Kay, G. K. Eaglesham, N. Holling, and P. A. Ryan, "Investigation of cuticular hydrocarbons for determining the age and survivorship of Australasian mosquitoes," *American Journal of Tropical Medicine and Hygiene*, vol. 74, pp. 462–474, mar 2006.

[294] K. H. S. Peiris, B. S. Drolet, L. W. Cohnstaedt, and F. E. Dowell, "Infrared Absorption Characteristics of Culicoides sonorensis in Relation to Insect Age," *American Journal of Agricultural Science and Technology*, vol. 2, no. 2, pp. 49–61, 2014.

[295] R. Connelly, "Highlights of Medical Entomology 2018: The Importance of Sustainable Surveillance of Vectors and Vector-Borne Pathogens," *Journal of Medical Entomology*, vol. 56, pp. 1183–1187, sep 2019.

[296] B. J. Cassone, K. Mouline, M. W. Hahn, B. J. White, M. Pombi, F. Simard, C. Costantini, and N. J. Besansky, "Differential gene expression in incipient species of Anopheles gambiae," *Molecular Ecology*, vol. 17, no. 10, pp. 2491–2504, 2008.

[297] L. Vannini, T. W. Reed, and J. H. Willis, "Temporal and spatial expression of cuticular proteins of Anopheles gambiae implicated in insecticide resistance or differentiation of M/S incipient species," *Parasites and Vectors*, vol. 7, p. 24, jan 2014.

[298] K. R. Reidenbach, C. Cheng, F. Liu, C. Liu, N. J. Besansky, and Z. Syed, "Cuticular differences associated with aridity acclimation in African malaria vectors carrying alternative arrangements of inversion 2La," *Parasites and Vectors*, vol. 7, pp. 1–13, apr 2014.

[299] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, pp. 1–8, feb 2006.

[300] L. C. Lee, C.-Y. Y. Liong, and A. A. Jemain, "A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum," *Chemometrics and Intelligent Laboratory Systems*, vol. 163, pp. 64–75, apr 2017.

[301] N. B. Gallagher, T. A. Blake, and P. L. Gassman, "Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil," *Journal of Chemometrics*, vol. 19, pp. 271–281, may 2005.

[302] M. Zeaiter and D. Rutledge, "Preprocessing Methods," in *Comprehensive Chemometrics*, vol. 3, pp. 121–231, Elsevier, jan 2009.

[303] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common preprocessing techniques for near-infrared spectra," 2009.

[304] A. Savitzky and M. J. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, pp. 1627–1639, jul 1964.

[305] M. Maleki, A. Mouazen, H. Ramon, and J. De Baerdemaeker, "Multiplicative scatter correction during on-line measurement with near infrared spectroscopy," *Biosystems Engineering*, vol. 96, no. 3, pp. 427–433, 2007.

[306] R. Barnes, M. Dhanoa, and S. Lister, "Correction to the description of standard normal variate (snv) and de-trend (dt) transformations in practical spectroscopy with applications in food and beverage analysis—2nd edition," *NIR news*, vol. 5, no. 3, pp. 6–6, 1994.

[307] Q. Guo, W. Wu, and D. L. Massart, "The robust normal variate transform for pattern recognition with near-infrared data," *Analytica Chimica Acta*, vol. 382, pp. 87–103, feb 1999.

[308] H. Wold, "Systems under indirect observation using pls," *A second generation of multivariate analysis: Methods*, 1982.

[309] M. Bevilacqua, R. Bucci, A. D. Magrì, A. L. Magrì, R. Nescatelli, and F. Marini, "Classification and Class-Modelling," in *Data Handling in Science and Technology*, vol. 28, pp. 171–233, Elsevier, jan 2013.

[310] D. J. Siria, R. Sanou, J. Mitton, E. P. Mwanga, A. Niang, I. Sare, P. C. D. Johnson, G. Foster, A. M. G. Belem, K. Wynne, R. Murray-Smith, H. M. Ferguson, M. González-Jiménez, S. A. Babayan, A. Diabaté, F. O. Okumu, and F. Baldini, "Rapid ageing and species identification of natural mosquitoes for malaria surveillance," *bioRxiv*, p. 2020.06.11.144253, jan 2020.

[311] S. Rajpurohit, R. Hanus, V. Vrkoslav, E. L. Behrman, A. O. Bergland, D. Petrov, J. Cvačka, and P. S. Schmidt, "Adaptive dynamics of cuticular hydrocarbons in Drosophila," *Journal of Evolutionary Biology*, vol. 30, pp. 66–80, jan 2017.

[312] S. Rajpurohit, V. Vrkoslav, R. Hanus, A. G. Gibbs, J. Cvačka, and P. S. Schmidt, "Post-eclosion temperature effects on insect cuticular hydrocarbon profiles," *Ecology and Evolution*, vol. 11, pp. 352–364, jan 2021.

[313] F. Menzel, B. B. Blaimer, and T. Schmitt, "How do cuticular hydrocarbons evolve? Physiological constraints and climatic and biotic selection pressures act on a complex functional trait," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, mar 2017.

[314] K. M. Wagoner, T. Lehmann, D. L. Huestis, B. M. Ehrmann, N. B. Cech, and G. Wasserberg, "Identification of morphological and chemical markers of dry- and wet-season conditions in female Anopheles gambiae mosquitoes," *Parasites and Vectors*, vol. 7, pp. 1–13, jun 2014.

[315] A. Gloria-Soria, J. Soghigian, D. Kellner, and J. R. Powell, "Genetic diversity of laboratory strains and implications for research: The case of Aedes aegypti," *PLoS Neglected Tropical Diseases*, vol. 13, p. e0007930, dec 2019.

[316] E. B. Ibitoye, I. H. Lokman, M. N. Hezmee, Y. M. Goh, A. B. Zuki, and A. A. Jimoh, "Extraction and physicochemical characterization of chitin and chitosan isolated from house cricket," *Biomedical Materials (Bristol)*, vol. 13, no. 2, p. 25009, 2018.

[317] E. Z. Panagou, F. R. Mohareb, A. A. Argyri, C. M. Bessant, and G. J. E. Nychas, "A comparison of artificial neural networks and partial least squares modelling for the rapid detection of the microbial spoilage of beef fillets based on Fourier transform infrared spectral fingerprints," *Food Microbiology*, vol. 28, pp. 782–790, jun 2011.

[318] L. Estelles-Lopez, A. Ropodi, D. Pavlidis, J. Fotopoulou, C. Gkousari, A. Peyrodie, E. Panagou, G. J. Nychas, and F. Mohareb, "An automated ranking platform for machine learning regression models for meat spoilage prediction using multi-spectral imaging and metabolic profiling," *Food Research International*, vol. 99, pp. 206–215, sep 2017.

[319] M. P. Milali, M. T. Sikulu-Lord, S. S. Kiware, F. E. Dowell, R. J. Povinelli, and G. F. Corliss, "Do NIR spectra collected from laboratory-reared mosquitoes differ from those collected from wild mosquitoes?," *PLoS ONE*, vol. 13, p. e0198245, may 2018.

[320] J. Perez-Mendoza, J. E. Throne, F. E. Dowell, and J. E. Baker, "Chronological age-grading of three species of stored-product beetles by using near-infrared spectroscopy," *Journal of Economic Entomology*, vol. 97, no. 3, pp. 1159–1167, 2004.

[321] C. Singh, D. Jayas, J. Paliwal, and N. White, "Detection of insect-damaged wheat kernels using near-infrared hyperspectral imaging," *Journal of stored products research*, vol. 45, no. 3, pp. 151–158, 2009.

[322] P. R. Griffiths, "Introduction to the Theory and Instrumentation for Vibrational Spectroscopy," in *Handbook of Vibrational Spectroscopy*, American Cancer Society, nov 2010.

[323] L. M. Ng and R. Simmons, "Infrared spectroscopy," *Analytical Chemistry*, vol. 71, jun 1999.

[324] M. Brandstetter, A. Genner, K. Anic, and B. Lendl, "Tunable Mid-IR lasers: A new avenue to robust and versatile physical chemosensors," *Procedia Engineering*, vol. 5, pp. 1001–1004, jan 2010.

[325] A. Ogunleke, V. Bobroff, H. H. Chen, J. Rowlette, M. Delugin, B. Recur, Y. Hwu, and C. Petibois, "Fourier-transform vs. quantum-cascade-laser infrared microscopes for histopathology: From lab to hospital?," *TrAC - Trends in Analytical Chemistry*, vol. 89, pp. 190–196, 2017.

[326] M. Erfan, Y. M. Sabry, B. Mortada, K. Sharaf, and D. Khalil, "Mid infrared MEMS FTIR spectrometer," in *MOEMS and Miniaturized Systems XV*, vol. 9760, p. 97600K, SPIE, mar 2016.

[327] D. W. Schiering and J. T. Stein, "Design considerations for portable mid-infrared ftir spectrometers used for in-field identifications of threat materials," *Portable Spectroscopy and Spectrometry*, pp. 41–65, 2021.

[328] A. O. Ghoname, Y. M. Sabry, M. Anwar, and D. Khalil, "Attenuated total reflection (atr) mems ftir spectrometer," in *MOEMS and Miniaturized Systems XIX*, vol. 11293, pp. 170–175, SPIE, 2020.

[329] M. Razeghi, W. Zhou, S. Slivken, Q.-Y. Lu, D. Wu, and R. McClintock, "Recent progress of quantum cascade laser research from 3 to 12 μm at the Center for Quantum Devices [Invited]," *Applied Optics*, vol. 56, no. 31, p. H30, 2017.

[330] M. Razeghi, Q. Y. Lu, N. Bandyopadhyay, W. Zhou, D. Heydari, Y. Bai, and S. Slivken, "Quantum cascade lasers: from tool to product," *Optics Express*, vol. 23, no. 7, p. 8462, 2015.

[331] B. Tuzson, M. Graf, J. Ravelid, P. Scheidegger, A. Kupferschmid, H. Looser, R. Paulo Morales, and L. Emmenegger, "A compact QCL spectrometer for mobile, high-precision methane sensing aboard drones," *Atmospheric Measurement Techniques*, vol. 13, pp. 4715–4726, sep 2020.

[332] D. Mammez, R. Vallon, B. Parvitte, M. H. Mammez, M. Carras, and V. Zéninari, "Development of an external cavity quantum cascade laser spectrometer at 7.5 μm for gas detection," *Applied Physics B: Lasers and Optics*, vol. 116, pp. 951–958, feb 2014.

[333] A. Schwaighofer, M. R. Alcaráz, C. Araman, H. Goicoechea, and B. Lendl, "External cavity-quantum cascade laser infrared spectroscopy for secondary structure analysis of proteins at low concentrations," *Scientific Reports*, vol. 6, p. 33556, dec 2016.

[334] B. Bird and J. Rowlette, "High definition infrared chemical imaging of colorectal tissue using a Spero QCL microscope," *Analyst*, vol. 142, pp. 1381–1386, apr 2017.

[335] A. Schwaighofer, C. K. Akhgar, and B. Lendl, "Broadband laser-based mid-IR spectroscopy for analysis of proteins and monitoring of enzyme activity," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 253, p. 119563, may 2021.

[336] A. Schwaighofer, M. R. Alcaraz, L. Lux, and B. Lendl, "pH titration of β-lactoglobulin monitored by laser-based Mid-IR transmission spectroscopy coupled to chemometric analysis," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 226, feb 2020.

[337] M. R. Alcaráz, A. Schwaighofer, H. Goicoechea, and B. Lendl, "EC-QCL mid-IR transmission spectroscopy for monitoring dynamic changes of protein secondary structure in aqueous solution on the example of β-aggregation in alcohol-denaturated α-chymotrypsin," *Analytical and Bioanalytical Chemistry*, vol. 408, pp. 3933–3941, jun 2016.

[338] B. E. Brumfield and M. C. Phillips, "Application of rapidly-swept external cavity quantum cascade lasers for open-path and standoff chemical sensing," in *Micro- and Nanotechnology Sensors, Systems, and Applications X*, vol. 10639, p. 1063928, International Society for Optics and Photonics, may 2018.

[339] R. Ostendorf, L. Butschek, S. Hugger, F. Fuchs, Q. Yang, J. Jarvis, C. Schilling, M. Rattunde, A. Merten, J. Grahmann, D. Boskovic, T. Tybussek, K. Rieblinger, and J. Wagner, "Recent Advances and Applications of External Cavity-QCLs towards Hyperspectral Imaging for Standoff Detection and Real-Time Spectroscopic Sensing of Chemicals," *Photonics*, vol. 3, no. 4, p. 28, 2016.

[340] S. Welzel, F. Hempel, M. Hübner, N. Lang, P. B. Davies, and J. Röpcke, "Quantum cascade laser absorption spectroscopy as a plasma diagnostic tool: An overview," *Sensors*, vol. 10, pp. 6861–6900, jul 2010.

[341] J. Röpcke, P. B. Davies, N. Lang, A. Rousseau, and S. Welzel, "Applications of quantum cascade lasers in plasma diagnostics: A review," *Journal of Physics D: Applied Physics*, vol. 45, p. 423001, oct 2012.

[342] C. Ye, *Tunable external cavity diode lasers*. World Scientific, 2004.

[343] A. Schwaighofer, M. Montemurro, S. Freitag, C. Kristament, M. J. Culzoni, and B. Lendl, "Beyond Fourier Transform Infrared Spectroscopy: External Cavity Quantum Cascade Laser-Based Mid-infrared Transmission Spectroscopy of Proteins in the Amide i and Amide II Region," *Analytical Chemistry*, vol. 90, no. 11, pp. 7072–7079, 2018.

[344] T. Skov, F. Van Den Berg, G. Tomasi, and R. Bro, "Automated alignment of chromatographic data," *Journal of Chemometrics*, vol. 20, pp. 484–497, nov 2006.

[345] M. R. Jung, F. D. Horgen, S. V. Orski, V. Rodriguez C., K. L. Beers, G. H. Balazs, T. T. Jones, T. M. Work, K. C. Brignac, S. J. Royer, K. D. Hyrenbach, B. A. Jensen, and J. M. Lynch, "Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms," *Marine Pollution Bulletin*, vol. 127, pp. 704–716, feb 2018.

[346] Z. Chen, J. N. Hay, and M. J. Jenkins, "The thermal analysis of poly(ethylene terephthalate) by FTIR spectroscopy," *Thermochimica Acta*, vol. 552, pp. 123–130, jan 2013.

[347] K. M. Elkins, *Introduction to forensic chemistry*. CRC Press, 2018.

[348] D. N. Ingebrigtson and A. Lee Smith, "Infrared Analysis of Solids by Potassium Bromide Pellet Technique," *Analytical Chemistry*, vol. 26, no. 11, pp. 1765–1768, 1954.

[349] R. H. Wilson, A. C. Smith, M. Kacurakova, P. K. Saunders, N. Wellner, and K. W. Waldron, "The mechanical properties and molecular dynamics of plant cell wall polysaccha-

rides studied by Fourier-transform infrared spectroscopy," *Plant Physiology*, vol. 124, no. 1, pp. 397–405, 2000.

[350] J. M. Kriesel, C. N. Makarem, M. C. Phillips, J. J. Moran, M. L. Coleman, L. E. Christensen, and J. F. Kelly, "Versatile, ultra-low sample volume gas analyzer using a rapid, broad-tuning ECQCL and a hollow fiber gas cell," in *Next-Generation Spectroscopic Technologies X*, vol. 10210, p. 1021003, SPIE, may 2017.

[351] X. Jia, L. Wang, Z. Jia, N. Zhuo, J. Zhang, S. Zhai, J. Liu, S. Liu, F. Liu, and Z. Wang, "Fast Swept-Wavelength, Low Threshold-Current, Continuous-Wave External Cavity Quantum Cascade Laser," *Nanoscale Research Letters*, vol. 13, p. 341, dec 2018.

[352] A. Lyakh, R. Barron-Jimenez, I. Dunayevskiy, R. Go, E. Tsvid, and C. K. N. Patel, "Progress in rapidly-tunable external cavity quantum cascade lasers with a frequency-shifted feedback," apr 2016.

[353] S. Pengel, B. Schönberger, S. Nayak, and A. Erbe, "Attenuated total reflection mid-IR-spectroscopy for electrochemical applications using a QCL," in *Laser Applications to Chemical, Security and Environmental Analysis, LACSEA 2012*, 2012.

[354] Y. Matsuura, K. Yoshioka, and S. Kino, "Blood glucose measurement with multiple quantum cascade lasers using hollow-optical fiber-based ATR spectroscopy," in *Optical Fibers and Sensors for Medical Diagnostics and Treatment Applications XVIII* (I. Gannot, ed.), vol. 10488, p. 12, SPIE, feb 2018.

[355] T. Koyama, S. Kino, T. Sasaki, Y. Wada, R. Kasahara, Y. Oba, and Y. Matsuura, "Measurement and uniformization of power distribution on the prism for biomedical applications of mid-infrared, attenuated-total-reflection spectroscopy," in *Optical Fibers and Sensors for Medical Diagnostics, Treatment and Environmental Applications XXI*, vol. 11635, p. 41, SPIE, mar 2021.

[356] N. J. Galán-Freyle, L. C. Pacheco-Londoño, A. D. Román-Ospino, and S. P. Hernandez-Rivera, "Applications of quantum cascade laser spectroscopy in the analysis of pharmaceutical formulations," *Applied Spectroscopy*, vol. 70, no. 9, pp. 1511–1519, 2016.

[357] R. Müller, M. Haertelt, J. Niemasz, K. Schwarz, V. Daumer, Y. V. Flores, R. Ostendorf, and R. Rehm, "Thermoelectrically-cooled inas/gasb type-ii superlattice detectors as an alternative to hgcdte in a real-time mid-infrared backscattering spectroscopy system," *Micromachines*, vol. 11, no. 12, pp. 1–13, 2020.

[358] W. C. Lin and S. J. Matcher, "Swept-source OCT using pulsed mid-infrared light," in *Label-free Biomedical Imaging and Sensing (LBIS) 2019* (N. T. Shaked and O. Hayden, eds.), vol. 10890, p. 84, SPIE, mar 2019.

[359] M. Brandstetter, C. Koch, A. Genner, and B. Lendl, "Measures for optimizing pulsed EC-QC laser spectroscopy of liquids and application to multi-analyte blood analysis," in *Quantum Sensing and Nanophotonic Devices XI* (M. Razeghi, E. Tournié, and G. J. Brown, eds.), vol. 8993, p. 89931U, SPIE, dec 2013.

[360] P. C. D. Hobbs, "Shot noise limited optical measurements at baseband with noisy lasers," in *Laser Noise*, vol. 1376, pp. 216–221, SPIE, mar 1991.

[361] Thorlabs, "Galvo-Resonant Scanners and Controllers," 2021.

[362] S. M. R. Motaghian Nezam, "High-speed polygon-scanner-based wavelength-swept laser source in the telescope-less configurations with application in optical coherence tomography," *Optics Letters*, vol. 33, p. 1741, aug 2008.

[363] J. Cao, P. Wang, Y. Zhang, G. Shi, B. Wu, S. Zhang, and Y. Liu, "Methods to improve the performance of the swept source at 10 μm based on a polygon scanner," *Photonics Research*, vol. 5, p. 245, jun 2017.

[364] A. Yoshihara, A. Miyazaki, T. Maeda, Y. Imai, and T. Itoh, "Spectroscopic characterization of dragonfly wings common in Japan," *Vibrational Spectroscopy*, vol. 61, pp. 85–93, jul 2012.

[365] E. R. Lucas, A. C. Darby, S. J. Torr, and M. J. Donnelly, "A gene expression panel for estimating age in males and females of the sleeping sickness vector Glossina morsitans," *PLOS Neglected Tropical Diseases*, vol. 15, no. 9, pp. 1–15, 2021.

[366] D. Mahajan, R. Girshick, V. Ramanathan, M. Paluri, and L. van der Maaten, "Advancing state-of-the-art image recognition with deep learning on hashtags," 2018.

[367] F. R. Giorgetta, E. Baumann, M. Graf, Q. Yang, C. Manz, K. Köhler, H. E. Beere, D. A. Ritchie, E. Linfield, A. G. Davies, Y. Fedoryshyn, H. Jäckel, M. Fischer, J. Faist, and D. Hofstetter, "Quantum cascade detectors," *IEEE Journal of Quantum Electronics*, vol. 45, no. 8, pp. 1039–1052, 2009.

[368] R. Szedlak, A. Harrer, M. Holzbauer, B. Schwarz, J. P. Waclawek, D. Macfarland, T. Zederbauer, H. Detz, A. M. Andrews, W. Schrenk, B. Lendl, and G. Strasser, "Remote Sensing with Commutable Monolithic Laser and Detector," *ACS Photonics*, vol. 3, pp. 1794–1798, oct 2016.

[369] J. Hillbrand, L. Matthieu Krüger, S. Dal Cin, H. Knötig, J. Heidrich, A. Maxwell Andrews, G. Strasser, U. Keller, and B. Schwarz, "High-speed quantum cascade detector characterized with a mid-infrared femtosecond oscillator," *Optics Express*, vol. 29, p. 5774, feb 2021.

[370] A. Dabrowska, M. David, S. Freitag, A. M. Andrews, G. Strasser, B. Hinkov, A. Schwaighofer, and B. Lendl, "Broadband laser-based mid-infrared spectroscopy em-

ploying a quantum cascade detector for milk protein analysis," *Sensors and Actuators B: Chemical*, vol. 350, p. 130873, jan 2022.

[371] J. Grahmann, A. Merten, R. Ostendorf, M. Fontenot, D. Bleh, H. Schenk, and H.-J. Wagner, "Tunable External Cavity Quantum Cascade Lasers (EC-QCL): an application field for MOEMS based scanning gratings," in *MOEMS and Miniaturized Systems XIII*, vol. 8977, p. 897708, SPIE, mar 2014.