Using Machine Learning and Infrared Spectroscopy for Rapid Assessment of Key Entomological Indicators of Malaria Transmission

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

by

Emmanuel Peter Mwanga

School of Biodiversity, One Health, and Veterinary Medicine
College of Medical, Veterinary & Life Sciences

University of Glasgow

Contents

List of Tables

List of Figures

Definition of key terms

Infrared spectroscopy: refers to a scientific method that uses light waves, specifically in the infrared range, to study and identify the chemical composition of substances. It works by shining infrared light on a material and measuring how the light is absorbed or reflected, allowing the detection of unique "fingerprints," such as molecules in the samples.

Mid-infrared spectroscopy: is a type of infrared spectroscopy that focuses on light waves in the middle range of the infrared spectrum. This range is particularly useful for studying organic molecules, as it can reveal detailed information about their biochemical structure and composition.

Machine learning: is a branch of computer science that enables computers to learn and make decisions or predictions from data with minimal human intervention. This is achieved through algorithms that learn and improve over time by repeatedly processing data, a process known as training.

Transfer learning: refers to a machine learning technique where a model trained for one task is adapted to perform a similar but different task. It helps save time and resources by reusing knowledge from an existing model. For example, a model trained on mosquito data from one region can be adjusted to work in another region with different mosquito populations.

Dimensionality reduction: is a method used to simplify complex data by reducing the number of features or variables. This process removes noise or redundant information while retaining the most important information in the data.

Malaria: is a disease caused by a parasites of the genus *Plasmodium* that infect humans through the bites of infectious female *Anopheles* mosquitoes. It is a significant public health concern in many tropical and subtropical regions.

Summary

Malaria vector surveillance is a critical element in control and elimination programs in endemic regions, serving to assess current transmission levels, vector species behaviours, and the efficacy of control interventions. Key surveillance metrics typically include the density and diversity of biting *Anopheles* mosquitoes, their blood-feeding histories, parasite prevalence within vectors, and the age structure of adult mosquito populations, among other indicators. However, conventional methods for monitoring these metrics are often costly, labour-intensive, and time-consuming, underscoring the need for scalable, simple, and cost-effective alternatives. The work presented in this thesis aligns with the recommendations of key policy organisations, including the World Health Organisation (WHO), which advocate for integrating effective surveillance into malaria control strategies in endemic regions.

The primary aim of my PhD project was to demonstrate that the emerging approach of Mid-infrared spectroscopy combined with machine learning (MIRS-ML) – a method that analyses biochemical signals generated by infrared light absorption in a sample – can offer high-throughput, and accurate assessments of entomological and parasitological indicators of malaria transmission. The project was therefore designed to provide field validation for the application of this technology by addressing several critical gaps to facilitate the effective implementation of MIRS-ML in vector surveillance. These gaps included: 1) the necessity for field-calibrated models to predict key entomological indicators of malaria across diverse settings, 2) the need to demonstrate the efficacy of this approach in areas where *Anopheles funestus* is predominant, as this species is the most significant malaria vector in East and Southern Africa but had not been analysed using MIRS-ML, 3) the need to apply this approach to multiple indicators in both laboratory and field settings, and 4) the necessity to show that infectious mosquitoes harbouring *Plasmodium* sporozoites in their salivary glands can be reliably detected using MIRS-ML.

The specific objectives of my PhD thesis were therefore as follows: 1) To evaluate the usefulness of transfer learning and dimensionality reduction techniques for improving the generalisability and transferability of MIRS-ML-based predictions for mosquito age classifications, 2) To demonstrate the application of MIRS-ML in classifying epidemiologically relevant age categories of adult female *An. funestus* mosquitoes, 3) To

demonstrate the field applicability of MIRS-ML for identifying blood meal sources in field-collected *An. funestus* mosquitoes, 4) To validate the field applicability of MIRS-ML for detecting *Plasmodium*-infected *An. funestus* mosquitoes, and 5) To explore key lessons learned from infrared-based entomological and parasitological studies, and to outline future directions for the use of MIRS-ML in malaria surveillance. The field studies were conducted in an area in Southeastern Tanzania where *An. funestus* accounts for more than 80% of malaria transmission.

In objective 1 (Chapter 2), I explored whether dimensionality reduction and transfer learning could improve the generalisability of MIRS-based age predictions. Here, the dimensionality of the spectra data was reduced using unsupervised principal component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE), and then used to train deep learning and standard ML models. Transfer learning was also used to reduce computational costs and enhance generalisability when predicting mosquito ages from new populations. The findings indicated that while dimensionality reduction alone did not improve generalisability, it did reduce computational time. Transfer learning was crucial for achieving generalisable MIRS-ML models for mosquito age prediction, suggesting that combining it with dimensionality reduction can improve the efficiency, transferability, and dissemination of these models.

In objective 2 (Chapter 3), I focused on applying MIRS-ML to rapidly classify the epidemiologically relevant age categories of *An. funestus*. Spectra data were divided into two age categories: 1-9 days (young, non-infectious) and 10-16 days (old, potentially infectious). PCA was used to reduce dimensionality, and a set of standard ML models and multi-layer perceptron (MLP) were trained to predict mosquito age categories. The results demonstrated the effectiveness of MIRS-ML in quickly classifying epidemiologically relevant age groups of *An. funestus*. Having been previously applied to *Anopheles gambiae*, *Anopheles arabiensis* and *Anopheles coluzzii*, this demonstration on *An. funestus* supports the potential of this low-cost, reagent-free technique for widespread use across all major Afro-tropical malaria vectors.

In objective 3 (Chapter 4), I demonstrated the first field application of MIRS-ML for assessing the blood-feeding histories of malaria vectors, with direct comparison to polymerase chain reaction (PCR) assays. After scanning mosquito samples on a spectrometer, blood meals were confirmed by PCR to establish the 'ground truth' for training ML models. Logistic regression and MLP models achieved over 88% accuracy in predicting mosquito blood meal sources, as well as closely matching the human blood index (HBI) estimates with the PCR-based standard HBI. This chapter provided evidence for the utility of MIRS-ML as a complementary surveillance tool in settings where conventional molecular techniques are impractical, given its cost-effectiveness, simplicity, scalability, along with its generalisability, outweighing minor gaps in HBI estimation.

In objective 4 (Chapter 5), I demonstrated the first field application of MIRS-ML for rapid and accurate detection of *Plasmodium* sporozoite in wild-caught *An. funestus* mosquitoes without requiring laboratory reagents. Desiccated mosquito head and thoraxes were scanned on MIRS, and sporozoite infection were confirmed by enzyme-linked immunosorbent assay (ELISA) and PCR, to establish references for training ML models. The ML models accurately predicted sporozoite-infectious mosquito samples with ~92% classification accuracy, highlighting the potential of MIRS-ML to enhance surveillance in malaria-endemic regions.

Building on the findings from objective 1, 2, 3 & 4, chapter 6 discusses key lessons learned from infrared-based entomological and parasitological studies and explored the future prospects for MIRS-ML in malaria surveillance. While significant advances have been made, challenges such as improving model generalisability across different environments and enhancing the interpretability of biochemical signals remain. Transfer learning can improve model performance, but no single approach fully address the variability of field samples. The broader implementation of MIRS-ML for malaria surveillance will require continuous data generation, model validation, and the development of deployment-ready systems, including the potential use of pooled samples as current scanning is limited to individuals.

In conclusion, this thesis demonstrates the potential of MIRS-ML for reagent-free assessments of key entomological indicators of malaria transmission in field settings including mosquito age, blood-feeding histories, and *Plasmodium* infections. Another important advancement was the successful application of transfer learning and dimensionality reduction to improve model generalisability and computational efficiency across different mosquito populations. MIRS-ML achieved high accuracy in classifying epidemiologically relevant age groups, detecting blood meal sources, and identifying sporozoite-infected mosquitoes. While challenges such as data variability and model robustness remain, this research highlights the potential of the MIRS-ML approach as a powerful, reagent-free alternative to traditional surveillance methods. Future work should focus on optimising model performance and developing deployment-ready systems for multi-variable assessments in real-world settings, particularly in resource-limited, malaria-endemic regions.

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.

Acknowledgements

Completing this PhD is one of my greatest achievements. I have never celebrated any of my academic accomplishments before, but this one is special, and I shall celebrate it. During my time as a PhD candidate, I had the flexibility to spend time in both Tanzania and Glasgow, UK. This change of environment was crucial in recharging my energy and motivation to successfully complete my research. The success of this project was possible due to the continuous and immense support from my supervisors (Dr. Simon A. Babayan, Prof. Fredros O. Okumu, and Dr. Francesco Baldini), colleagues at the University of Glasgow and Ifakara Health Institute, and my family.

Foremost, I want to express my sincere gratitude to Simon Babayan for being a wonderful supervisor. I have learned many lessons from you in my academic career, including coding, statistics, and machine learning which made my project very smooth. Simply put, I learned from the best, and the knowledge I acquired will benefit me and others for many years to come. Moreover, you are an incredible human being and always considerate. I am delighted to have worked with you. You have shown incredible trust in me since the journey began with my master's fellowship.

I will forever be grateful to Francesco Baldini. You are one of the best, always like a father to his son. Thank you very much for your mentorship in vector and parasite biology. Your constant push and ambitions have driven my inner passion as a researcher. You spent plenty of time in the laboratory with me to ensure I learned from the best. Thank you very much for trusting me from the very beginning and agreeing to be one of my supervisors and mentors. This goes all the way back to the start of my master's fellowship in 2018 to the present.

Fredros Okumu, my friend, my deepest gratitude to you. You have been a friend, a brother and a supervisor. I will forever be grateful for the mentorship that goes beyond that of a PhD supervisor. You met me when I was as a young man, but the trust you put in me is immeasurable. When we met, you asked me what I wanted to do, and I replied, "I want to make an impact through research". You then told me, "Let's go to Ifakara, and you will make an impact". Since then, you have tirelessly trained me to become a competent researcher. The day I was applying for my Wellcome trust fellowship, you spent 2 days

with me in the lab, day and night, until we clicked the submission button, I can't imagine how many people would do that. I will forever be grateful. Thank you very much.

My sincere thanks to Prof. Samson Kiware, who has been a true mentor in my professional career and social life. Samson, you are always just a call away, and your support has been beyond explanation. Every step I have taken in my research career started with you, because without you I wouldn't have met Fredros. Additionally, the social support you have provided to me and my young family is something for which we are forever grateful.

I would like to extend my gratitude to my colleagues at the Ifakara Health Institute for their immense support and contributions toward the completion of this PhD. Special thanks to all members of the Outdoor Mosquito Control group (OMC) for their logistical and financial support in every aspect. Many thanks to colleagues and all members of the Vectors and Malaria group in Glasgow, from whom I have benefited greatly.

I am endlessly grateful to my parents, Mr. and Mrs. Peterson Salewi Mwanga, for their continuous support and encouragement, both socially and spiritually, throughout my entire PhD period.

I am grateful to the Wellcome Trust fellowship, Gates Foundation, and the Medical Research Council (MRC) for the financial support. I also acknowledge the American Society of Tropical Medicine (ASTMH) for the travel award that allowed me to attend the international annual meeting to share my research findings.

Finally, to my wife Thecla Muganyizi and our two boys, Axel and Harvey. Thecla, you have been unbelievable and a strong partner during my absence. Thank you for your wonderful support during my PhD journey. I know at times you and the kids still needed my time. I owe you all my time. Thank you very much for your moral support, love and prayers during this time. Indeed, I felt encouraged when I was down. *Gracias, Mi Amor*.

List of Abbreviations

ATR-FTIR	Attenuated total reflectance Fourier-Transform Infrared spectrometer
BMI	Blood meal index
CNN	Convolutional neural network
DL	Deep learning
ELISA	Enzyme-linked immunosorbent assay
ITNs	Insecticide treated nets
MIRS	Mid-infrared spectroscopy
MIRS-ML	Mid-infrared spectroscopy and machine learning
NIRS	Near-Infrared spectroscopy
PCA	Principal component analysis
PCR	Polymerase chain reaction
RDT	Rapid diagnostic test
SNP	Single nucleotide polymorphism
t-SNE	t-distributed stochastic neighbour embedding

Chapter 1

General Background

Mosquito-borne diseases impose a significant burden in Africa, with mosquitoes being among the world's greatest contributors to death and misery. They transmit numerous viruses and parasites to both humans and animals. The most important mosquito-borne diseases affecting humans include malaria, dengue, chikungunya, yellow fever, and lymphatic filariasis, among others. Many of these diseases are endemic across the region, and some occur frequently as outbreaks across all continents rural and urban areas. Given the tight linkages with environment, many of these diseases are expected to worsen due to climate change factors, notably warming temperatures and increased frequency of floods [1].

Malaria stands as the most significant and well-documented of mosquito-borne illnesses, with 249 million cases and 608,000 deaths reported globally in 2022 [2]. The majority of these cases and deaths are concentrated in sub-Saharan African countries, accounting for 94% and 96% of global malaria-related cases and deaths, respectively [2]. Other important mosquito-borne diseases include dengue, responsible for millions of cases and thousands of deaths annually though data on its prevalence are limited compared to malaria [3,4], and lymphatic filariasis, a parasitic nematode infection that is on the verge of elimination in many settings [5].

Malaria control is one of the current priority initiatives in sub-Saharan African countries and has been earmarked for elimination by various national governments and the African Union. Despite notable progress since 2000, there has also been significant stagnation since ~2015, with many high-burden countries reporting increased malaria cases in recent years [6]. It is estimated that in the WHO African region, malaria control efforts have reduced cases by 82% and deaths by 94% since 2000 [2]. Yet, the disease remains one of the leading public health challenges, with ~580,000 malaria deaths in WHO-Africa alone [2]. Several countries have established strategies to eliminate the disease within the next 10-20 years, all of them planning on the use of multiple strategies. For example, in Tanzania, malaria prevalence in children has significantly dropped from 14% in 2008 to 8% in 2022 [7,8].

This reduction is partly due to large-scale implementation of vector control strategies, such as insecticide treated nets (ITNs) and indoor residual spraying (IRS) [9,10]. There is a strategy (National Malaria strategic Plan 2021-2025: Transitioning to Malaria Elimination in Phases) aimed at further reducing malaria prevalence to less than 1% by 2025 [11]. However, achieving this target remains far-fetched due to persistent malaria transmission.

ITNs are mosquito nets treated with insecticide that repel or kill adult mosquitoes upon contact. ITNs are the main vector control strategy in Africa and have been estimated to contribute to two-thirds of all gains in malaria control since 2000 [12]. Another key strategy is IRS, which involves spraying the interior surface of houses with residual insecticides to kill or repel indoor-resting mosquitoes. IRS has been widely used for most of the past century and was at the core of the first Global Malaria Eradication campaign [13]. However, the effectiveness of these tools is currently threatened by various challenges, notably the rise of insecticide resistance and changes in mosquito behaviour from biting indoors to outdoors during early hours before bedtime [14–16]. This shift in behaviour creates a protection gap because mosquitoes can now evade the protective coverage of bed nets by biting individuals before they are under the nets.

1.1 Malaria parasite and life cycle

Malaria is caused by protozoan parasites belonging to the genus *Plasmodium*, which complete their life cycle in vertebrates and *Anopheles* mosquitoes, referred to as vectors (Fig. 1.1). Five species of *Plasmodium* are known to infect humans: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *P. knowlesi* – the latter primarily infecting monkeys but occasionally causing infections in humans [17,18]. The life cycle of the *Plasmodium* parasite is divided into three phases: the sporogonic cycle in the mosquito, and two phases in the human host – the erythrocytic cycle (within red blood cells) and exo-erythrocytic cycle (outside red blood cells) [19].

When a mosquito bites an infected person, it ingests gametocytes along with the blood [17]. The blood itself is necessary for the development of the mosquito's eggs. In the mosquito's gut, the gametocytes that have matured into male and female gametes fuse to form a zygote. The zygote transforms into an ookinete, which penetrates the gut wall and develops into an oocyst. Within the oocyst, the nucleus divides repeatedly, producing many sporozoites. These sporozoites are then released when the oocyst bursts, and they migrate to the mosquito's salivary glands, ready to be transmitted to a new human host during the next blood meal. The entire sporogonic cycle typically takes about 8-15 days, depending on temperature and species [20].

The infected *Anopheles* mosquito then bites a human and injects the sporozoite into the bloodstream. These sporozoites travel to the liver, where they multiply. After 7-12 days, the liver schizonts bursts, releasing merozoite into the bloodstream, where they invade red blood cells and continue asexual replication cycle [17]. Some of parasites develop into gametocytes, which are then ingested by another *Anopheles* mosquito, continuing the cycle [17].

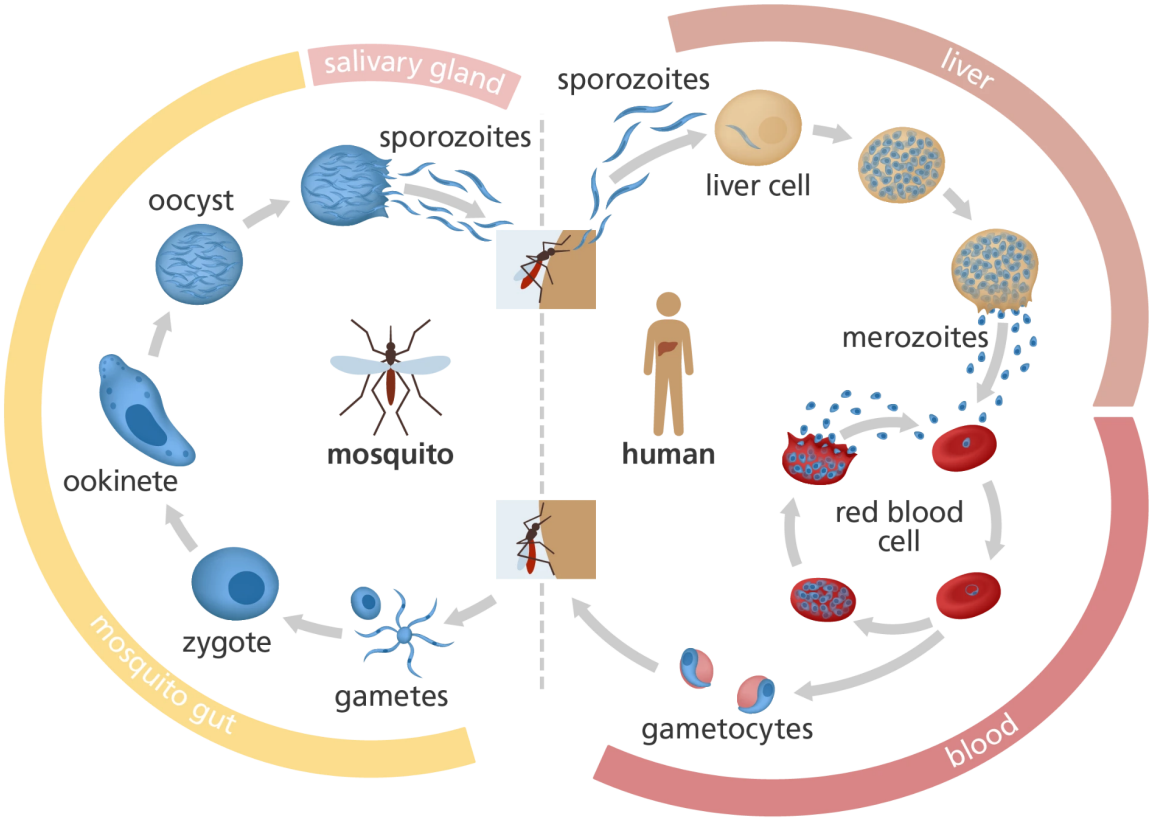

Figure 1.1: Illustration of the malaria Parasite life cycle. Image credit: Laura Olivares Boldú / Wellcome Connecting Science.

1.2 The life cycle of *Anopheles* mosquito

The life of an *Anopheles* mosquitoes consists of four stages: egg, larva, pupa and adult (Fig. 1.2). After a blood-meal, female mosquitoes develop eggs, which they lay after 3-4 days [19]. The choice of oviposition sites varies, with mosquitoes ovipositing eggs in locations such as small and large pools, streams, swamps, rivers, ponds, lakes, rice fields, or containers near human dwellings [19,21–23].

The eggs hatch into larvae within 1-3 days, and these larvae feed on organic particles in the water. Under normal tropical conditions, the larvae stage lasts about 8-10 days, although cooler temperatures can extend this period [19]. The larvae then transform into a non-feeding pupa, which undergoes metamorphosis and emerges as a flying adult within 1-3 days. Mating occurs soon after emergence, and females seek out blood meals, repeating the cycle. The duration of each developmental stage is influenced by environmental factors such as temperature and nutrition, with faster development occurring in warmer conditions. Some species, like *Anopheles funestus*, develop more slowly [24].

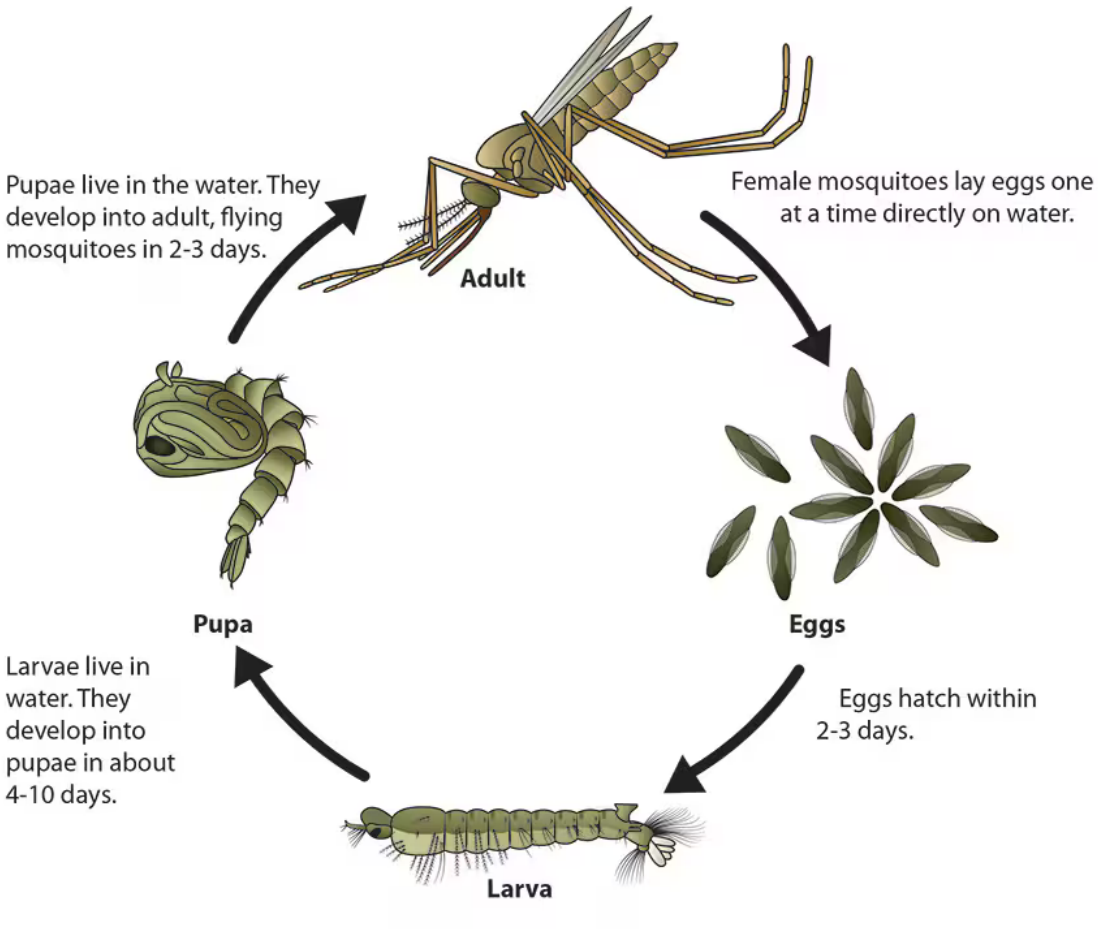

Figure 1.2: Illustration of the life cycle of *Anopheles*. Image credit: CDC

There are around 30 species of *Anopheles* mosquitoes that are of major importance in transmitting malaria [25]. In Africa, the major Afro-tropical malaria vectors include *Anopheles gambiae sensu stricto*, *Anopheles arabiensis*, *Anopheles coluzzii* and *An. funestus* [26–28]. Additionally, there is also the rise of invasive vector species, notably *Anopheles stephensi*, which is currently spreading in eastern Africa and has potential to increase the risk of malaria in urban settings [29–31].

1.3 Surveillance of malaria vectors and transmission

Vector surveillance is an essential component of malaria control and elimination programs in Africa for assessing prevailing transmission intensities, the behaviours of different vector species, for planning key interventions and assessing their responsiveness to different interventions [32]. The key metrics typically include the biting densities of various vector species, insecticide resistance status, proportion of mosquitoes that have blood-fed on humans or other vertebrates, prevalence of malaria parasites in the vector species, densities

of the immature stages, and age distribution of the adult mosquito populations, among others. The most important and commonly used metric for estimating malaria transmission is Entomological Inoculation rate (EIR), which quantifies the number of infectious bites per person over a given period of time. It is calculated as the product of the human biting rate (HBR) and proportion of the biting mosquitoes carrying *Plasmodium* sporozoite in their salivary glands [33–35]. The metric is typically used to estimate exposure levels and evaluate the effectiveness of control programs; and is one of the most direct measures of disease transmission. EIR is additive and is calculated for different species then summed up. It can also be calculated for indoor and outdoor biting vectors separately and then summed up. Typically, it is estimated annually but can also be scaled down to monthly or daily approximate EIR estimates.

Studies have shown that while there are statistical correlations between EIR and the epidemiological burden of malaria in different settings [36], these statistical correlations do not always hold across all transmission intensities. They are especially spurious in low transmission settings, where the statistical correlation is either completely lost or significantly weakened [37]. At high transmission intensities, it is also common for values to saturate, meaning additional increases in malaria transmission intensities do not necessarily lead to higher malaria prevalence. This situation is particularly common in areas where significant populations of people have developed some form of natural immunity to the disease [38]. Therefore, while EIR is one of the easiest and most widely used measures of malaria transmission, caution is required when using the data to inform intervention strategies. In settings nearing malaria elimination, EIR estimates can become difficult to obtain, as common methods for entomological estimation become poorly sensitive. For example, in one study in Ifakara, Tanzania, an area where long periods of ITNs coverage, gradual urbanisation with improved housing, and improved health systems has led to significant declines in disease, researchers had to sample continuously for over 3500 trap nights using different methods to obtain just one malaria-infected *Anopheles* [39]. Without this single infected mosquito, estimating the EIR would have been impossible. In such settings, it has been suggested to either measure the receptivity of an area (based on presence or absence of competent vector species and importation of the parasites) [40] or simply categorise EIR as either above 1 infectious bites/per/year (ib/p/y) (elimination unlikely) or below 1 ib/p/y (elimination possible).

Another important measure of malaria transmission risk is the proportion of all blood meals mosquitoes take from humans compared to other vertebrate blood meals, commonly referred to as human blood index (HBI), as malaria is not zoonotic (with exception of *P. knowlesi*). HBI typically measures the propensity of mosquitoes to feed on humans and thus transmit pathogens to them. Though increasingly ignored, it is one of the strongest indicators of malaria vector competence and has been considered an important indicator of malaria transmission in different settings [41]. This metric is reported to be particularly

high in major Afro-tropical malaria vectors, including *An. gambiae*, *An. funestus* and *An. coluzzii*, which are well-adapted to human environments [27]. The high anthropophily – i.e., the preference to bite humans – of these vectors is one of the main reasons malaria control in Africa has been lagging [27].

To estimate the HBI, mosquitoes must be collected from multiple sites in ways that are agnostic of their blood-feeding, resting or host seeking behaviours, covering both indoor and outdoor locations in human dwellings. The abdominal content of the females is then analysed by either enzyme-linked immunosorbent assay (ELISA) or polymerase chain reaction (PCR) to determine the main vertebrate blood sources. Metagenomic sequencing can improve understanding of the range of host species that mosquitoes depend on in the area [42]. Mosquitoes that primarily bite humans are more likely to be involved in pathogen transmission, while those that regularly bite both humans and animals may transmit zoonotic infections [18,43–49]. Previous studies have shown that additional blood meals post-infection is also an important factor in the ability of the vector species to transmit the disease. The growth and maturity of parasite stages, particularly oocysts, significantly increase with additional blood meals [50,51].

Thirdly, one can estimate mosquito age and survivorship, using these data to assess transmission risk, the overall demographic characteristics of vector populations, and the performance of interventions. Age-grading of mosquitoes is particularly important for determining the likelihood that mosquitoes will live long enough to allow complete parasite development (the extrinsic incubation period), and subsequent transmission to humans [52]. For malaria infections, it is expected that vector populations constantly targeted by ITNs and IRS or other adult-targeting interventions are likely to have far younger populations than those without such interventions. This assessment is essential for monitoring the impact of interventions like ITNs and IRS, which primarily target adult mosquitoes in the field [53].

Overall, accurate determination of the age, blood-feeding histories, and infectious status of malaria vector species are important indicators of their feeding behaviour, role in ongoing malaria transmission, and the overall risk exposure of people within those settings. However, measuring these indicators at intervention sites remains challenging, necessitating scalable, simple-to-implement, and low-cost methods for quantification. The work contained in this thesis was aimed at addressing these issues, and to allow better scalability of measuring entomological markers of transmission at low cost. In part, this effort responds to the WHO recommendation that countries incorporate effective surveillance as a core malaria control strategy and also the desire of the malaria endemic countries for low-cost approaches. The following section provides more detailed information on how these measures are currently conducted, and the advantages and disadvantages of the different approaches.

1.4 Mosquito age classification

Accurate estimation of mosquito age and survival probabilities is crucial for monitoring transmission dynamics and assessing the impact of vector control interventions. This metric is particularly important for malaria, as typically only a small subset of adult mosquitoes survive long enough to transmit the disease. The *Plasmodium* parasite requires approximately 10-12 days to mature from the time the male and female gametes are ingested by the mosquito to the time mature sporozoites are ready in the salivary glands [20]. The development of both the parasite and the vector depends on climatic factors and can be accelerated in warmer temperatures. In much of tropical Africa, it is possible to estimate and assign thresholds for chronological age (number of days lived) and biological age (referring to physiological maturity of mosquitoes), beyond which parasite transmission can occur. Biological age is particularly important when considered alongside other physiological developments such as blood-meal acquisition and exposure of susceptible humans.

The primary entomological technique for estimating mosquito age is the dissection of mosquito ovaries [53]. This process involves using microscopy to examine the reproductive history of mosquitoes. When dissected, the ovaries are inspected for coiled tracheolar skeins, which indicate non-parous (young) mosquitoes, or stretched-out tracheole, which indicate parous (older) mosquitoes that may carry malaria parasites due to multiple blood-feedings [53]. Dissection can also be used to determine how many times a female mosquito has previously laid eggs by counting the tracheolar constrictions [54]. This method has also been used for assessing dwarf ovarioles, which typically do not mature into full egg production, to estimate the number of times the mosquito has previously been gravid; while mostly reliable, it is difficult to obtain approximation of the population-level age of the mosquitoes using these methods.

A major challenge associated with this approach is that these dissection methods are not scalable, given that they are labour-intensive and time-consuming [53,54]. Additionally, the reliability of dissection is compromised by the reproductive history of mosquitoes. For instance, gonotrophic discordance, where a female mosquito takes multiple blood meal without laying eggs, can skew the result [55]. Additionally, when trying to estimate the number of completed gonotrophic cycles, mosquitoes that have recently laid eggs may still be having an open ovariolar sac, making it impossible to assess the number of previous gravidity events. Thirdly, relevant in the context of malaria transmission, which requires an incubation period of 9-14 days [20], differentiating between distinct categories of adult mosquitoes based solely on parity is not always informative. Therefore, there is a need for alternative age-grading techniques that are cost-effective, scalable, and capable of providing an accurate representation of mosquito age categories and populations.

Alternative mosquito age classification methods have been explored in different settings, though mostly on a small scale in research settings. One example is transcriptional profiling, which involves analysing the expression of age-related genes [56]. This molecular approach can offer high precision in age determination, although it requires sophisticated laboratory equipment and expertise. Another method investigated is age-grading through the examination of cuticle protein degradation [57]. This method relies on identifying specific protein markers that degrade over time, offering insights into the chronological age of the mosquito. Another innovative approach is the use of wing scales on the fringes, where the degradation of these scales can be used as a marker of mosquito age [58]. More recently, there has been growing interest in the use of infrared spectroscopy, which utilises the absorption or reflectance of specific wavelengths of light to estimate the age of mosquitoes by analysing their cuticular hydrocarbons. This technique has shown promise in differentiating between young and old mosquitoes, providing a non-destructive and rapid assessment method [59–61], as described in more detail below. These methods, although promising, have yet to be validated in large-scale field studies. In summary, the transition from experimental to practical, large-scale application will require addressing the need for standardisation, affordability, and lack of specialised training in laboratory techniques.

1.5 Identification of vertebrate blood meal sources to understand mosquito blood-feeding histories

Identifying mosquito blood meal sources is crucial for understanding host-vector interactions and providing important information on the transmission dynamics of important vector-borne diseases, including malaria, lymphatic filariasis, Zika, dengue, Japanese encephalitis, Rift Valley fever, Chikungunya and West Nile virus, pose significant risks to humans [5, 17, 18, 43–49]. Blood-feeding by mosquitoes is important for two main reasons: for the mosquito itself, it is essential for egg production, reproductive fitness, and at times serves as a source of metabolic energy; it is also important for the pathogens which mosquitoes transmit, as it enables them to complete their life cycle in a new host [62, 63]. Additionally, evidence suggests that the host selection pattern of anthropophagic mosquitoes may be influenced by factors such as the presence of vector-control strategies like ITNs, which provides a physical barrier and have a killing effect, reducing mosquito exposure to humans [64]. This may lead mosquitoes to seek alternative hosts, particularly livestock [64], which can have a zoopotentiation effect. This effect refers to an increased tendency of mosquitoes to feed on humans who live near livestock [65, 66], as the livestock emit heat and odour cues that attract mosquitoes. Consequently, zoophagic mosquitoes find additional blood sources, and even naturally

anthropophagic mosquitoes may feed on cattle when host cues are mixed nearby. This can increase disease transmission risk by providing alternative blood meal sources, consequently increasing mosquito survival rates and abundance [67].

Host blood meal sources are primarily identified using a range of tools that have evolved over time, from immunological assays like ELISA [68] to nucleic acid assays such PCR [69]. Both techniques offer a reasonable degree of specificity and sensitivity, enabling accurate identification of host species from even small amounts of blood, to distinguish between different blood meal types based on the vertebrate antigens and deoxyribonucleic acid (DNA). ELISA uses host-specific antibody-enzyme conjugate to detect homologous immunoglobulin G (IgG) in blood-fed mosquito samples, thereby indicating the host species [68]. In contrast, PCR assays rely on the extracted DNA from the sample to target the mitochondrial cytochrome b (cytB) protein, the protein encoded by the mitochondrial genome [70,71]. Since cytB has a high copy number as a mitochondrial gene, it is effective in identifying mosquito blood meals [69]. These techniques are now commonly used in many laboratories for identification of the source of arthropod blood meals, thanks to the commercial availability of antibody-enzyme conjugate for ELISA and primers to amplify homologous DNA fragments.

Proper collection and handling of mosquito samples are crucial for the accurate identification of blood meal sources. Mosquitoes can be collected using various traps, including aspirators, light traps, or mechanical aspirators. Host-baited traps are usually inappropriate because they target host-seeking mosquitoes which are typically not yet blood-fed; instead, the focus is primarily on resting mosquitoes that are already engorged. To ensure the best results, it is important to preserve the sample properly to prevent the degradation of host DNA or proteins in the blood meal. Specimens should be stored in tubes containing silica gel desiccant or frozen at -20°C or lower until analysis. For molecular assays, preserving samples in ethanol or RNAlater can help maintain DNA integrity and improve the quality of the tests [72,73]. Additionally, care should be taken to avoid cross-contamination between samples during collection and processing.

The ELISA and PCR approaches for identifying mosquito blood meals each have their advantages and disadvantages. Immunological assays, such as ELISA, are relatively straightforward, cost-effective, and can be performed with minimal equipment, making them suitable for field use. However, they are also prone to cross-reactivities when testing multiple hosts, or poor detection thresholds when the samples have not been appropriately handled, potentially leading to false negatives [74]. PCR-based methods offer higher sensitivity and specificity, and can identify blood meals from mixed host sources [69]. These methods also allow for detection of non-host pathogens within the blood meal. However, they require more sophisticated laboratory infrastructure and are more expensive. Where there is adequate capacity, these nucleic acid-based tests can be

expanded to include metagenomic analysis of mosquito gut content [42]. This approach allows for a comprehensive representation of the full repertoire of vertebrate hosts that the mosquitoes have previously bitten. By utilizing metagenomic techniques, researchers can identify a wide array of host DNA present in the mosquito gut, providing a more detailed understanding of host-vector interactions and enhancing surveillance of vector-borne disease transmission dynamics [42].

1.6 Detection of infective *Plasmodium* sporozoite adult *Anopheles* mosquitoes

Detection of infective *Plasmodium* sporozoite in adult *Anopheles* mosquitoes is equally important, serving as a key parameter in estimating human exposure to infectious mosquitoes over time. Several techniques can be used to detect *Plasmodium* sporozoite infections in mosquitoes. These include techniques such as dissection of salivary glands using microscopy [75]. Loop-mediated isothermal Amplification (LAMP) assays [76] which enable DNA amplification under isothermal conditions [77]. However, in ELISA and PCR are the most common used techniques in malaria surveillance [75,78,79]. ELISA measures the presence of *Plasmodium* circumsporozoite (CS) protein originating from sporozoites and have been developed and standardised for human malaria parasite like *P. falciparum*, *P. vivax*, *P. malariae*, and *P. ovale* [75,80–83]. However, ELISA can yield false positives if non-target protozoans are present, potentially leading to an overestimate of the sporozoite prevalence and subsequent EIR [84,85]. Additionally, ELISA requires specific antibodies, repeated reagents, and well-trained personnel, making it both costly and time-consuming. For most local laboratories in malaria endemic countries, such supplies are not always readily available even if the unit prices are also high and controlled [86].

PCR-based methods offer higher specificity and sensitivity by amplifying *Plasmodium* DNA from mosquito samples, allowing for the detection of lower-density infections that ELISA might miss [78]. These molecular techniques can differentiate between species of *Plasmodium* in a single multiplexed reaction, providing detailed epidemiological data [78]. However, PCR assays also require significant laboratory infrastructure, including thermocyclers, electrophoresis equipment, and access to high-quality reagents, primers and probes, which may not be widely or readily available in field laboratories in endemic settings. Additionally, they necessitate highly trained personnel and stringent contamination controls, further increasing the operational cost and complexity. Despite these challenges, advances in high-throughput and field-deployable PCR systems, along with improved sample preservation techniques, are enhancing the feasibility of these methods for large-scale epidemiological and entomological surveillance.

1.7 Using infrared spectroscopy to analyse key entomological and parasitological indicators of malaria transmission

Advances in infrared spectroscopy have opened new possibilities for entomological and parasitological investigations into the transmission of mosquito-borne pathogens, including malaria. Infrared spectroscopy, encompassing near-infrared (NIR, 12,500 - 4,000 cm^{-1}), mid-infrared (MIR, 4,000 - 400 cm^{-1}), and far-infrared (FIR, 1,000 - 50 cm^{-1}) regions, can be utilized for various analytical purposes in biological sciences. NIR spectroscopy, for example, is often used for rapid and non-destructive analysis of water content and organic compounds [87,88], making it useful for studying plant-insect interactions and assessing the physiological state of insects [89,90]. It is also used in assessing the quality of food and pharmaceutical products [91] and, changes in age and infectivity of mosquitoes [60,92–95]. FIR spectroscopy, although less commonly used, provides detailed information on the low vibrational region of vibrational spectra, aiding in the study of inter – and intramolecular structures and dynamics within biological samples [96].

Other studies have also used Raman spectroscopy, which relies on the inelastic scattering of monochromatic light, usually from a laser [97]. When this light interacts with a molecule, most photons are elastically scattered (Rayleigh scattering), but a small fraction of light is scattered at different energies, corresponding to the vibrational modes of the molecule (Raman scattering) [97]. This technique is particularly advantageous for its ability to provide detailed molecular fingerprints without the need for extensive sample preparation [98]. It is widely used in various fields such as pharmaceuticals [99,100], earth and material science [101], and forensic analysis [102]) due to its high sensitivity and specificity. Most recently, surface-enhanced Raman spectroscopy has also been used for mosquito age-grading [103,104].

Visible wavelengths have also been used in different spectroscopic techniques, such as visible absorption spectroscopy and fluorescence spectroscopy [105, 106]. Visible absorption spectroscopy measures the absorption of visible light by a sample, providing information about the electronic transitions of molecules, which can be related to their chemical structure and concentration [105]. This method is often used in chemical analysis, quality control, and environmental monitoring and has also been applied in parasitological investigations of malaria infections [107–112]

Among these spectroscopy approaches, mid-infrared spectroscopy (MIRS) has shown remarkable promise for entomological and parasitological applications. MIRS records spectral information on the biochemical composition of samples within the 4,000 - 400 cm^{-1} frequency range (Fig. 1.3), providing structural identities of molecules present

through well-delineated absorption bands [97,98]. This technique measures key biological components such as lipids, proteins, and chitin [59], which can vary with mosquito age and species, the presence or absence of parasites, increased cuticle thickness in resistant mosquitoes, and different mosquito host blood meal sources [61,113–117].

Infrared techniques offer significant advantages, particularly because they are quick and do not require any reagents aside from desiccants for sample preservation and controlling humidity in the spectrometers. This makes them a potentially cost-effective technique for malaria surveillance in resource-limited areas. Further, the ability to analyse the biochemical composition of mosquito samples rapidly and accurately without the need for extensive sample preparation or chemical reagents is a crucial benefit for large-scale field applications. While the analysis of large amounts of infrared spectral data requires appropriate analytical techniques, the advancements in machine learning now make it possible to automate most of these processes and more efficiently analyse the voluminous mosquito and parasite data.

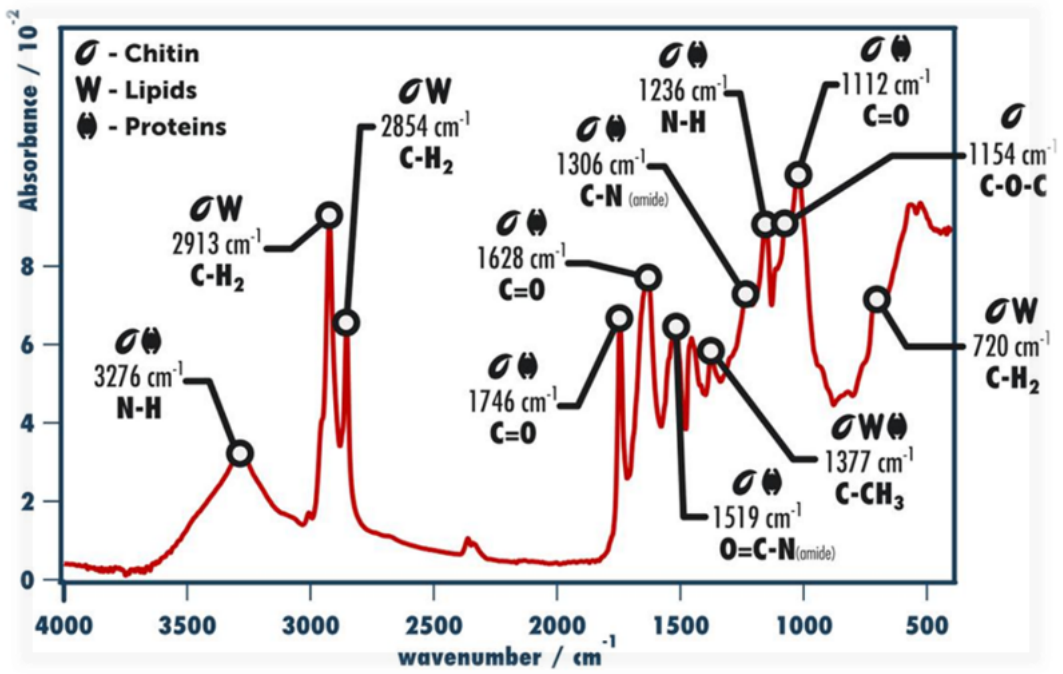

Figure 1.3: Assignment of spectral bands in MIR spectra showing different chemical composition of a mosquito sample. By relying on the fundamental molecular vibration of C-H, N-H, O-H and S-H functional groups, MIR produces distinctive spectra for closely similar molecules making them particularly invaluable in distinguishing biological samples. Image credit: Dr. Mario González-Jiménez.

1.8 Machine learning

Machine learning (ML) is a branch of computer science that enables computers to make decisions based on data with minimal human input [118]. This capability is achieved through algorithms that learn and improve over time by repeatedly processing data, a process known as training [119]. During training, the algorithm is provided with data inputs (features) and corresponding outputs (labels), allowing it to optimise and refine its decision-making ability. The model's performance is evaluated by comparing its prediction to the true values. Once trained, the model can make predictions or decisions from new, unseen data [118]. Ultimately, the goal is for the model to generalise well, meaning it can make accurate predictions across a variety of new, unseen data with no or minimal retraining. ML has broad applications across various scientific fields, including biomedical sciences, chemistry, and pattern recognition [120–122].

ML can be broadly categorised into supervised and unsupervised learning. Unsupervised learning does not rely on labelled data; instead, it identifies patterns or relationships within data and is often used for tasks like association, clustering, and dimensionality reduction. Dimensionality reduction simplifies data by reducing the number of features (or variables), removing noise or redundant information while retaining the important information [119, 123]. Techniques like principal component analysis (PCA) are frequently used in fields like spectroscopy for visualising and analysing data, as well as for dimensionality reduction [124, 125].

In contrast, supervised learning uses labelled data, where the algorithm learns from specific input-output pairs [118]. After training, the algorithm is tested on new data to estimate its accuracy. If the results are not satisfactory, the training process is repeated until the model's performance improves. Supervised learning is often divided into classification, where items are sorted into categories, and regression, which predicts continuous values [118]. Popular supervised learning algorithms include K-nearest neighbours, linear regression, logistic regression, support vector machines, decision trees, random forests, and neural networks. Additionally, other learning methods, such as semi-supervised learning, reinforcement learning, and deep learning, offer alternative approaches to solving complex problems [118, 119]. Algorithms are typically chosen based on their suitability for the task at hand, the size and complexity of the dataset, and the desired outcome.

Pre-trained or newly trained ML algorithms can then be used to identify patterns and correlations in spectral data that may not be apparent through traditional analytical methods, thereby enhancing the accuracy and speed of entomological assessments using MIRS.

1.9 Research focus and objectives

The core focus of this PhD thesis is to investigate the potential of MIRS combined with different ML techniques to measure various entomological and parasitological parameters typically assessed during malaria vector and transmission surveillance. The work is motivated by the need to develop cost-effective techniques for rapidly and effectively estimating entomological and parasitological indicators of malaria at scale in resource-limited areas.

Several studies have already documented the potential of MIRS combined with ML in estimating key entomological and parasitological indicators of malaria transmission [59,61,113–115,126]. Most of these studies have been conducted in semi-field environments, with only few involving field-collected samples. For example, Siria *et al.*, used MIRS-ML to predict both age and species of *An. gambiae*, *An. arabiensis*, and *An. coluzzii* from single samples collected under laboratory and semi-field conditions [61].

A major challenge reported in this study was the lack of generalisability. The models often failed to accurately predict unseen data from different locations due to inherent variability in mosquitoes arising from differences in environmental conditions, mosquito populations, genetic factors, and dietary factors [61]. Moreover, previous research focused on a limited range of entomological indicators, primarily the identification of species and the determination of their age, either individually or as a population. To address these concerns, one part of this thesis aimed to investigate whether the generalisability of ML models can be improved by using transfer learning (i.e., updating a pre-trained model with a small amount of new data from a different target population) and reducing the dimensionality of the MIR spectra. The expectation was that this could improve the predictive accuracy of the MIRS-ML approach for mosquitoes from different locations.

Furthermore, while MIRS-ML has been successfully used to predict the age of three major Afro-tropical malaria vectors such as *An. gambiae*, *An. arabiensis*, and *An. coluzzii*, it was clear that the technique needed to be expanded to include other vectors, such *An. funestus*, whose role in malaria transmission has been increasing significantly [127–134]. This PhD thesis was therefore aimed also at broadening the application of MIRS-ML to effectively differentiate the epidemiologically relevant age of *An. funestus*.

In the preliminary studies that we conducted prior to the start of my PhD work, my colleagues and I had demonstrated that a MIRS-ML based-approach could distinguish and predict mosquito blood-feeding histories from four different hosts, achieving ~97% accuracy with spectra data recorded from the abdomens of laboratory-reared *An. arabiensis* [114]. We argued then that MIRS offered far better distinctive capabilities for the biological signals contained in the spectral data and showed that this system was a viable option for

multiple parameters. However, it was also noted that for MIRS-ML to be viable for malaria vector surveillance, field validation is crucial, as mosquitoes collected from the field comes from different ecological conditions that are likely to affect MIRS profiles. Moreover, MIRS-ML have not yet been used to estimate the infection status of field-collected malaria vectors. Therefore, this thesis evaluates the field performance of MIRS-ML based approaches, using molecular and immunological assays as the ground truth, to assess the host preferences of blood-feeding mosquitoes and their parasite infection status from field-collected *An. funestus*.

Finally, to ensure that MIRS-ML are effectively integrated into malaria surveillance, my thesis explores lessons learned from previous works and potential future directions.

1.10 Objectives

The overall objective of my thesis is to demonstrate that MIRS-ML based-approaches can provide high-throughput and accurate assessments of mosquito age, blood-feeding history, and detection of *P. falciparum* in field-collected mosquitoes. Additionally, it aims to evaluate the lessons learned and future directions for the wider use of MIRS-ML in malaria surveillance.

Specific objectives

1. Demonstrate the application of transfer learning and dimensionality reduction to MIRS data to improve the transferability and generalisability of MIRS-ML based predictions for mosquito age [CHAPTER 2].
2. Demonstrate the application of the MIRS-ML based approach to classify the epidemiologically relevant age of adult female *An. funestus* mosquitoes [CHAPTER 3].
3. Demonstrate the field application of MIRS-ML based approaches to detect blood meal sources of field-collected *An. funestus* [CHAPTER 4].
4. Demonstrate the field application of MIRS-ML based approaches to detect *Plasmodium*-infected *An. funestus* from field-collected samples. This objective focuses on *An. funestus*, which now contributes more than 80% of malaria transmission in Southeastern Tanzania [128] [CHAPTER 5].
5. Interrogate key lessons learned from infrared-based entomological and parasitological studies so far, and the potential future directions of the use of MIRS-ML in malaria surveillance [CHAPTER 6].

1.11 Geographical focus

The studies in this thesis were all done in Tanzania, where the primary malaria vectors are *An gambiae s.s*, *An. funestus*, and *An. arabiensis* [26, 28]. Among these species, *An. funestus* and *An. arabiensis* have become the main vectors of malaria transmission due to their high resistance to the insecticides commonly used in ITNs [128, 135]. As a result, the role of *An. funestus* in malaria transmission has increased significantly [28, 129, 132], now accounting for approximately 80% of ongoing malaria transmission, particularly in rural areas [128, 130].

Indeed, the increased prominence of *An. funestus* in malaria transmission is closely linked to its preference for human blood over animal blood [128, 136], and its adaptability to bite humans early before they go to bed [14, 15]. This behavioural plasticity, combined with its insecticide resistance, underscores the critical need for innovative surveillance and control strategies to address the growing challenge posed by *An. funestus* in malaria-endemic regions.

Chapter 2

Using Transfer Learning and Dimensionality Reduction Techniques to Improve Generalisability of Machine-learning Predictions of Mosquito Ages from Mid-infrared Spectra

Emmanuel P. Mwanga[1,2*], Doreen J. Siria[1,2], Joshua Mitton[2,3], Issa H. Mshani[1,2], Mario González-Jiménez[2,4], Prashanth Selvaraj[5], Klaas Wynne[4], Francesco Baldini[2,3], Fredros O. Okumu[1,2,6], Simon A. Babayan[2].

1. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, Morogoro, Tanzania
2. School of Biodiversity, One Health, and Veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, UK
3. School of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK
4. School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK
5. Institute for Disease Modelling, Bellevue, WA, 98005, USA
6. School of Public Health, University of Witwatersrand, Johannesburg, South-Africa

DOI: https://doi.org/10.1186/s12859-022-05128-5

2.1 Abstract

Background: Old mosquitoes are more likely to transmit malaria than young ones. Therefore, accurate prediction of mosquito population age can drastically improve the evaluation of mosquito-targeted interventions. However, standard methods for age-grading mosquitoes are laborious and costly. We have shown that Mid-infrared spectroscopy (MIRS) can be used to detect age-specific patterns in mosquito cuticles and thus can be used to train age-grading machine learning models. However, these models tend to transfer poorly across populations. Here, we investigate whether applying

dimensionality reduction and transfer learning to MIRS data can improve the transferability of MIRS-based predictions for mosquito ages.

Methods: We reared adults of the malaria vector *Anopheles arabiensis* in two insectaries. The heads and thoraces of female mosquitoes were scanned using an attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectrometer, which were grouped into two different age classes. The dimensionality of the spectra data was reduced using unsupervised principal component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE), and then used to train deep learning and standard machine learning classifiers. Transfer learning was also evaluated to improve the computational cost of the models when predicting mosquito age classes from new populations.

Results: Model accuracies for predicting the age of mosquitoes from the same population as the training samples reached 99% for deep learning and 92% for standard machine learning. However, these models did not generalise to a different population, achieving only 46% and 48% accuracy for deep learning and standard machine learning, respectively. Dimensionality reduction did not improve model generalisability but reduced computational time. Transfer learning by updating pre-trained models with 2% of mosquitoes from the alternate population improved performance to 98% accuracy for predicting mosquito age classes in the alternative population.

Conclusion: Combining dimensionality reduction and transfer learning can reduce computational costs and improve the transferability of both deep learning and standard machine learning models for predicting the age of mosquitoes. Future studies should investigate the optimal quantities and diversity of training data necessary for transfer learning and the implications for broader generalisability to unseen datasets.

Key terms: *Anopheles arabiensis*, convolutional neural network, standard machine learning, generalisability, dimensionality reduction, transfer learning.

2.2 Background

Malaria currently kills approximately one child every minute [137]. In 2020, there were 241 million cases and 627,000 deaths, nearly all in Sub-Saharan Africa [137]. Currently, the most widespread and cost-effective method of malaria prevention is based on controlling the mosquitoes that transmit the disease. Since 2000, insecticide-treated nets (ITNs) and indoor residual spraying (IRS) have so far contributed nearly 80% of all global malaria decline [12]. However, the direct impact of individual control programs on the mosquito populations and on malaria transmission at the sites of intervention remains difficult to measure. To guide further efforts against the disease, evaluating the performance of these

and other vector control interventions is crucial for measuring their impact in different settings. The World Health Organization (WHO) now recommends that surveillance be integrated as a core component of malaria control programs [32] .

This necessitates scalable, simple-to-implement and low-cost methods for quantifying key biological attributes of mosquitoes, such as age, infection status, and blood meal preferences, which are essential for understanding pathogen transmission dynamics. The age and survivorship of key *Anopheles* vectors are especially important in determining the likelihood that the mosquitoes will live long enough to allow complete parasite development (the extrinsic incubation period), and subsequent transmission to humans [52]. The assessments are essential for monitoring the impacts of interventions such as ITNs and IRS, which primarily kill adult mosquitoes in the field [53].

The current "gold standard" for estimating the age of malaria mosquitoes is to dissect their ovaries to estimate how many times they have laid eggs [53, 54]. Despite their low technical demands, such procedures are time-consuming and labour-intensive. Age-grading dissections can also be imprecise because of gonotrophic discordance, which is common in Afrotropical malaria vectors [55], or of their reliance on the availability of host blood meals, which determines when and how frequently a mosquito blood-feeds.

We and others have demonstrated that spectroscopic analysis of mosquitoes using near infrared (12,500 - 4,000 cm^{-1}) or mid-infrared (MIR) (4,000 - 400 cm^{-1}) frequencies can identify key biochemical signals that vary with age [59, 60]. These methods, when combined with specific machine learning (ML) techniques, allow for rapid estimation of mosquito ages [59, 61].

Despite early successes, these infrared-based applications have limitations such as their portability to mosquitoes from different locations or laboratories [61] and the substantial computational requirements for retraining such models. Indeed, the inherent variability of mosquitoes from different environmental and genetic backgrounds may limit the generalisability of models trained on infrared spectra. The models could also be misled by signals in MIRS that are associated with confounding factors introduced during sampling (e.g., atmospheric contamination with water vapour, temperature variations and high humidity in the laboratory), thus learning features that are not strictly related to the biochemical trait being investigated. Therefore, machine learning models must be regularly updated with new data from target mosquito populations.

To increase the generalisability of ML models for a given training dataset, a variety of spectral smoothing and regularisation techniques have been tested, such as penalised regression [138]. These methods are known to be computationally efficient and to improve generalisability [138]. Deep learning (DL) techniques such as convolutional neural networks (CNN) have recently been used on large spectra data [61], improving

generalisability through transfer learning (i.e., updating a pre-trained model with a small amount of new data from a different target population). However, when trained on large datasets, such techniques remain computationally expensive and may necessitate repeated sampling of hundreds of mosquitoes from different populations and environments to allow successful generalisability. Alternatively, since standard ML models are less complex than DL, computational time can be kept to a minimum. DL methods are versatile extensions of machine learning that are ideal for complex or large datasets [118]. But are prone to overfitting, such as predicting the training dataset well but failing on previously unseen or new data.

However, unsupervised learning algorithms, which find patterns independent of pre-defined target labels, can aggregate, cluster or eliminate features while retaining dominant statistical information before machine learning training on the spectra data. The resulting dimensionality reduction may improve generalisability, reducing overfitting, increasing the signal-to-noise ratio of the data, as well as lowering computational requirements for training machine learning models. Examples include principal component analysis (PCA) [124, 125, 139], which projects a large number of variables into distinct categories that summarise data into a small number of independent principal components, and t-distributed Stochastic Embedding (t-SNE) [140], which clusters data points based distances between all their input dimensions.

This study assessed whether the generalisability and computational costs of MIRS-based models for predicting the age classes of female *An. arabiensis* mosquitoes reared in two different insectaries in two locations could be improved by combining dimensionality reduction and transfer learning methods.

2.3 Methods

2.3.1 Collection of mosquito spectra data

We analysed mid-infrared spectra from two strains of *An. arabiensis* mosquitoes obtained from two different insectaries, one from University of Glasgow, UK and another from Ifakara Health Institute, Tanzania. The same data had previously been used to demonstrate the capabilities of mid-infrared spectroscopy and CNN for distinguishing between species and determining mosquito age [61]. The insectary conditions under which the mosquitoes were reared (temperature 27$\pm$1.0°C, and relative humidity 80$\pm$5%) have been described elsewhere [61].

Mosquitoes were collected from day 1 to day 17 after pupal emergence at both laboratories and divided in two age classes (1-9 day-olds and 10-17 day-olds). Silica gel was used to dry the mosquitoes. For each chronological age in each laboratory, 120 samples were measured by MIRS on each day. The heads and thoraces of the mosquitoes were then scanned with an attenuated total reflectance Fourier-Transform Infrared (FTIR) ALPHA II and Bruker Vertex 70 spectrometers both equipped with a diamond ATR accessory (BRUKER-OPTIC GmbH, Ettlingen, Germany). The scanning was performed in the mid-infrared spectral range (4,000 - 400 cm^{-1}) at a resolution of 2 cm^{-1}, with each sample being scanned 16 times to obtain averaged spectra as previously described [59, 114]. As a result, the spectral dataset contained 1665 spectral features (Fig. 2.1).

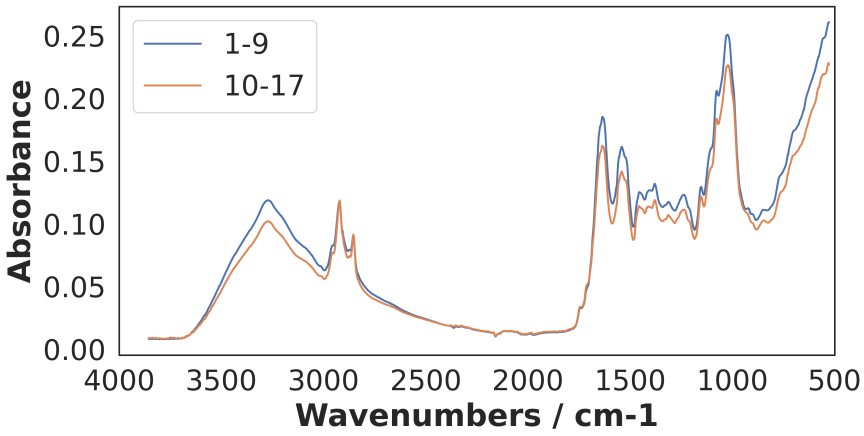

Figure 2.1: The Average mid-infrared spectra of dried mosquitoes aged 1-9 days and 10-17 days. The supervised learning was trained on the slight difference between mosquitoes aged 1-9 and 10-17 days

2.3.2 Data pre-processing

The spectral data were cleaned to eliminate bands of low intensity or significant atmospheric intrusion using the custom algorithm [59]. The final datasets from Ifakara and Glasgow contained 1,720 and 1,635 mosquito spectra, respectively. In these two datasets, the chronological age of *An. arabiensis* was categorised as 1-9 days old (i.e. young mosquitoes representative of those typically unable to transmit malaria) and 10-17 days old (i.e. older mosquitoes representative of those potentially able to transmit malaria) [20].

To improve the accuracy and speed of convergence of subsequent algorithms, data were standardised by centring around the mean and scaling to unit variance [141].

2.3.3 Dimensionality reduction

Principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) were used separately to reduce the dimensionality of the data [124, 125, 139, 140]. Both PCA and t-SNE were implemented using the scikit-learn library [141]. Separately, t-SNE was used to convert high-dimensional Euclidean distances between spectral points into joint probabilities representing similarities. To cluster the data into three features, the embedded space was set to 3, because the Barnes-hut algorithm in t-SNE is limited to only 4 components. Perplexity was set to 30 as the number of nearest neighbours, which means that for each point, the algorithm took the 30 closest points and preserved the distances between them. For smaller datasets perplexity values ranging from 5 and 50 are thought to be optimal for avoiding local variations and merged clusters caused by small or large perplexity values [140]. The learning rate for t-SNE is generally in the range of 10 - 1,000 [141], thus it was set to 200 scalar.

2.3.4 Machine learning training

Deep learning: DL models were trained and used to classify the *An. arabiensis* mosquitoes into the two age classes (1-9 or 10-17 day-olds). The intensities of *An. arabiensis* mid-infrared spectra (matrix of features) were used as input data, while the model outputs were the mosquito age classes.

Three different deep learning models were trained; 1) Convolutional neural network (CNN) model without dimensionality reduction, 2) Multi-Layer Perceptron (MLP) with PCA as dimensionality reduction, and 3) MLP with t-SNE as dimensionality reduction. For all models, a SoftMax layer was added to transform the non-normalized outputs of K-units in a fully connected layer into a probability distribution of belonging to either one of two age classes (1-9 or 10-17 days). Moreover, to compute the gradient of the networks, stochastic gradient boosting was used as an optimisation algorithm [142], and categorical cross-entropy loss was used for the classifier's metric.

To begin, we trained a one-dimensional CNN model with four convolutional layers and one fully connected layer when the dimensionality of the data was not reduced (Fig. 2.2A), and therefore consisting of 1,666 training features from the data. The one-dimensional CNN was used because it is effective at deriving features from fixed-lengths (i.e., the wavelengths of the mid-infrared spectra), and it has been previously been used efficiently with spectral data [61]. To extract features from spectral signals, the deep learning architecture used convolutional, max-pooled and fully connected layers. The convolutional operation was carried out with kernel sizes (window) of 8, 4, and 6, and a kernel window shift size

(stride) of either 1 or 2. For each kernel size, 16 filters were used to detect and derive features from the input data. Furthermore, given the size of the training data, the fully connected layer consisted of 50 neurons to reduce the model's complexity.

Moreover, batch normalisation layers were added to both models to improve model stability by keeping mean activation close to 0 and activation standard deviation close to 1. To reduce the likelihood of overfitting, dropout was used during model training to randomly and temporarily remove units from the network at a rate of 0.5 per step. Furthermore, after 50 rounds, early stopping was used to halt training when a validation loss stopped improving.

Dimensionality reduction: We trained two additional deep learning models, in this case Multi-Layer Perceptron (MLP), with PCA or t-SNE transformed input data (Fig. 2.2B). The models were trained with only fully connected layers ($n = 6$) containing 500 neurons each, given the limited number of training features to ensure performance and stability. To control for overfitting, the procedure was similar to that of the CNN above, except that early stopping was used to halt training when a validation loss stopped improving after 500 rounds.

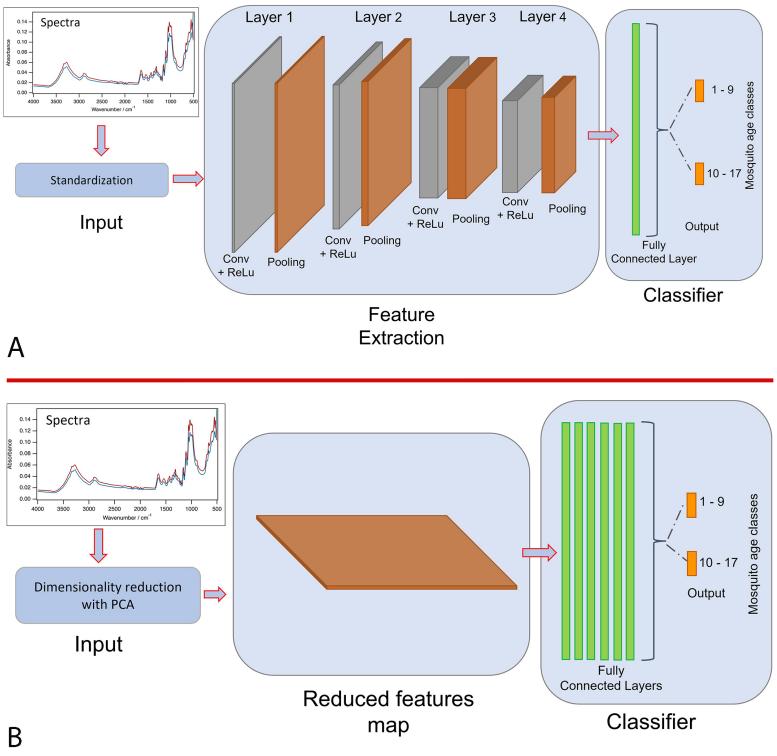

Figure 2.2: A schematic representation of a deep learning models that uses mosquito spectra as input to predict mosquito age classes. **A**) CNN - no dimensionality reduction is applied: standardised spectral features are fed as input through four different convolutional layers, followed by one fully connected layer, with the predicted age classes shown as the output layer. **B**) MLP - dimensionality reduction is used: spectral features that have been reduced in dimension using PCA or t-SNE are fed as input through 6 fully connected layers, with the predicted age classes shown as the output layer.

Transfer learning: The Ifakara dataset was used as the source domain for pre-training the ML model. The Ifakara dataset was divided into training and test sets, and estimator performance was assessed using K-fold cross-validation ($k = 5$) [143], (Fig. 2.3). We therefore determined what percentage of the new spectra data from the alternate location as target domain was required for ML models to learn the variability between the insectaries. To put transfer learning options to the test, either 82 or 33 spectra were randomly selected from the 1,635 of the Glasgow data, accounting for 5% and 2% of the dataset, respectively. The learning process in this case relied on a pre-trained model (trained with Ifakara data), avoiding the need to start training from scratch (Fig. 2.3). The ML models pre-trained with Ifakara dataset were fine-tuned using 2% or 5% subsets of the Glasgow dataset. The output was compared to that of a model trained solely with Ifakara data (i.e., no transfer learning).

Precision, recall, and F1-scores were calculated from predicted values for each age class to demonstrate the validity of the final models in predicting the unseen Glasgow data. Keras and TensorFlow version 2.0 were used for deep learning process [144, 145].

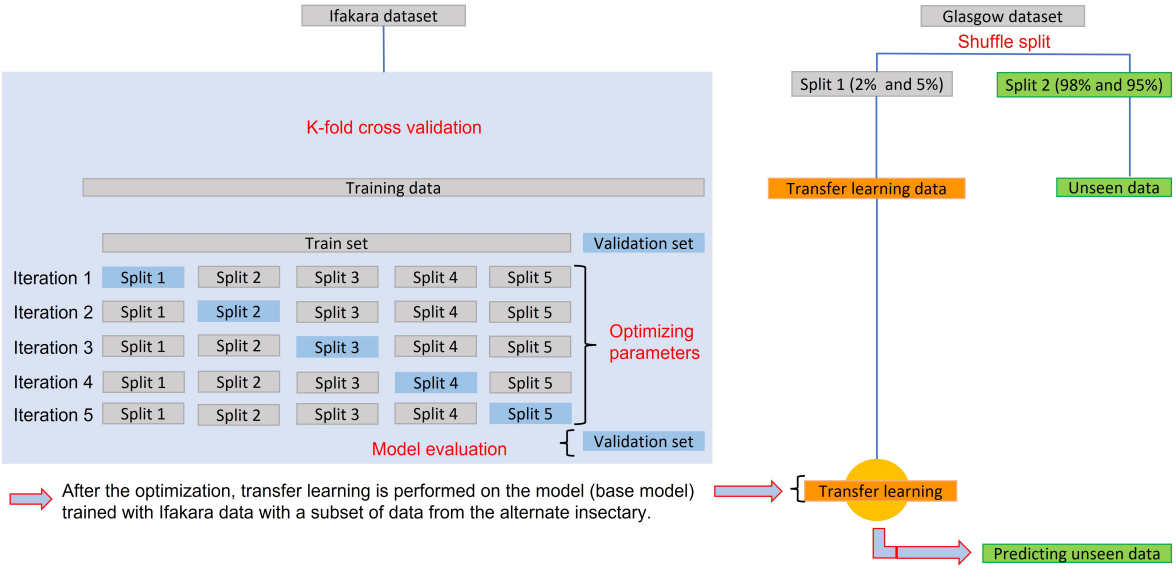

Figure 2.3: Schematic illustrating the process of data splitting, model training, cross-validation, and transfer learning.

Standard machine learning: We also compared the prediction accuracy of CNN to that of a standard machine learning model trained on spectra data transformed by PCA or t-SNE. Different algorithms were compared, including K-Nearest Neighbour, logistic regression, support vector machine classifier, random forest classifier, and a gradient boosting (XGBoost) classifier. The model with the highest accuracy score for predicting mosquito age classes was optimised further by tuning its hyper-parameters with randomised search cross-validation [141]. The cross-validation evaluation used to assess estimator performance in this case was the same as that used in deep learning. The

fine-tuned model was used to predict mosquito age classes in previously unseen Glasgow dataset.

Python version 3.8 was used for both the deep learning and standard machine learning training. All computations were done on a computer equipped with 32 Gigabytes of random-access memory (RAM) and an octa-core central processing unit. The 'best model' was the one that achieves high accuracy while maintaining low run times.

2.4 Results

2.4.1 Deep learning mosquito age classification with and without dimensionality reduction: Lack of generalisation between two locations

In the initial analysis, only spectra from the Ifakara insectary were used to train the CNN. During model training, the CNN classifier achieved 99% training accuracy without any dimensionality reduction (Fig. 2.4A). When given new held-out data from the same Ifakara insectary (test set), the model predicted mosquitoes aged 1-9 days with 100% accuracy and those aged 10-17 days with 99% accuracy (Fig. 2.4B). However, when the same model was used to predict age classes for Glasgow insectary samples, the overall accuracy was 46%, and therefore indistinguishable from any random classifications (Fig. 2.4C).

In addition, a CNN classifier required 200 epochs for training, with a running time of 7.2-7.8 seconds per epoch when no dimensionality reduction on the input data was used (Table 2.1).

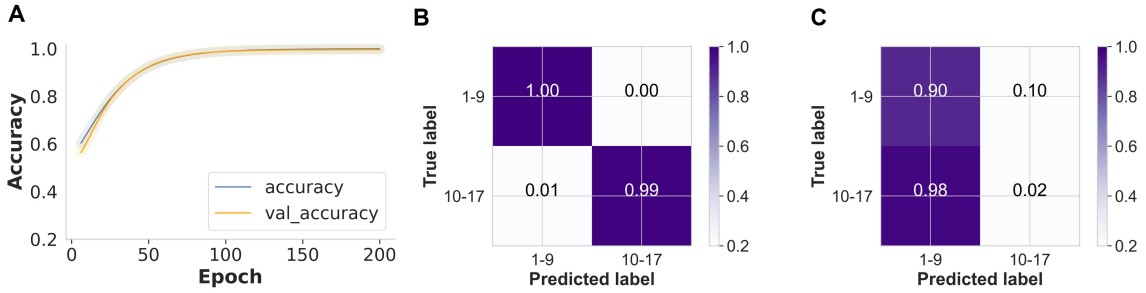

Figure 2.4: CNN generalisation and prediction of mosquito age using data from a single insectary (Ifakara) with no dimensionality reduction. **A**) Training and validation classification accuracy for mosquito age classes improved from 60% to 95% as training iterations increased (200 epochs). **B**) A normalised confusion matrix displaying the proportions of correct mosquito age class predictions achieved on the held-out Ifakara data (test set) during model training. **C**) Proportions of correct mosquito age class predictions based on unseen data from the alternate insectary (Glasgow).

PCA was used to project the data into lower dimensional space using singular value decomposition [125, 146], with the goal of achieving the best summary using optimal number of principal components (PCs) with up to 98% of variance explained (Fig. 2.5A). Further, when the impact of PCs on accuracy was assessed, a greater prediction accuracy was found, leading to the selection of 8 PCs (Fig. 2.5B).

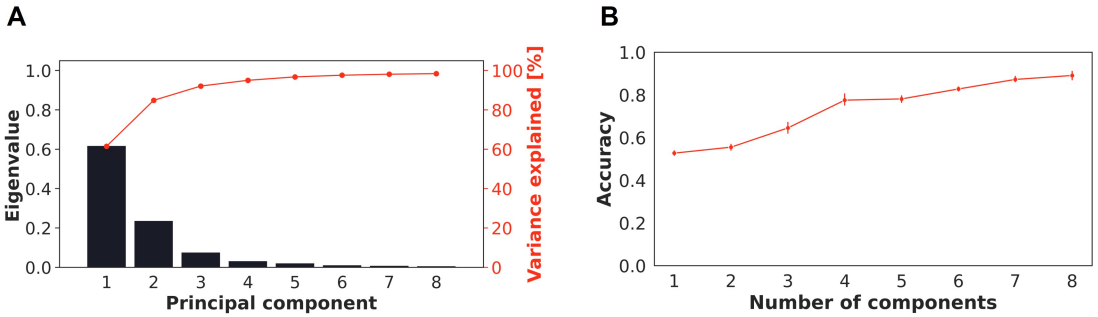

Figure 2.5: A) Cumulative explained variance and eigenvalues as the function of principal components. **B**) Number of principal components included in the XGB classifier (i.e. from 1:8 PCs)

When PCA was used to reduce the dimensionality of the data, the MLP model trained with only Ifakara spectra predicted the held-out data from the same insectary (Ifakara) with an overall accuracy of 91% but could attain only 58% accuracy for predicting age classes of Glasgow mosquitoes (Table 2.1). Similarly, when t-SNE was used as the dimensionality reduction technique, the model predicted the held-out Ifakara data (test set) with an accuracy of 85% but failed to accurately predict age classes of Glasgow data (Table 2.1).

Furthermore, when PCA or t-SNE were used to transform the input data, a MLP classifier needed 5,000 epochs to train, with a running time of 0.7-0.8 seconds per epoch (Table 2.1).

2.4.2 Transfer learning improves deep learning accuracy and generalisability

To improve generalisability (i.e., the ability of the models to predict the age classes of samples from other sources), we tuned the pre-trained CNN models with 2% or 5% of the spectra from Glasgow (i.e., 2% or 5% target population samples for transfer learning) and used the updated model to predict the unseen Glasgow dataset. When no dimensionality reduction was used, the pre-trained model predicted the held-out test (Ifakara dataset) with 99% accuracy and transferred well to the Glasgow dataset when 2% and 5% target population samples were used for transfer learning, achieving 100% and 96% accuracies, respectively (Table 2.1).

However, when PCA or t-SNE were used to reduce the dimensionality of the data, the MLP classifier was trained with only fully connected layers in this case to allow the model to learn the combination of features with the network's learnable weights. Using PCA, the pre-trained model predicted the held-out test (Ifakara dataset) with 91% accuracy, but when 2% transfer learning was applied, the model transferred well to the Glasgow dataset, achieving 97% accuracy when predicting the mosquito age classes, and 96% accuracy with 5% target population samples ((Table 2.1), Fig. 2.6A-C).

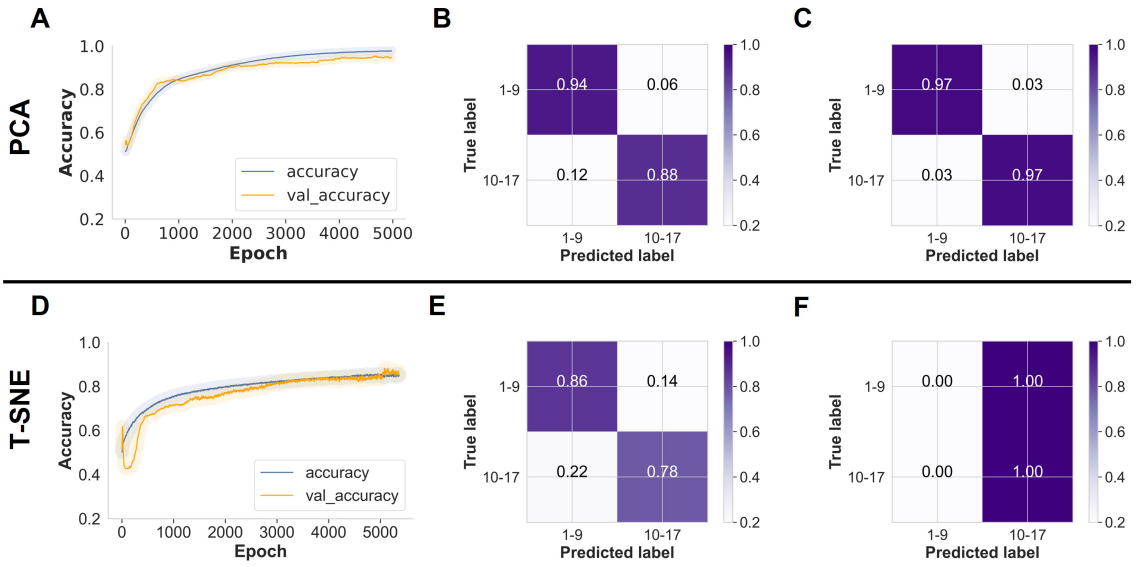

Figure 2.6: MLP trained on PCA-transformed Ifakara dataset plus 2% new target population samples: **A**) As training time increased (5,000 epochs), training and validation classification accuracy for mosquito age classes increased from 50% to 91%, **B**) A normalised confusion matrix displaying the proportions of correct mosquito age class predictions achieved on the held-out Ifakara test set during model training, **C**) Proportions of correct mosquito age class predictions achieved on unseen Glasgow dataset. MLP trained on t-SNE-transformed Ifakara dataset plus 2% new target population samples: **D**) As training time increased (5,000 epochs), training and validation classification accuracy for mosquito age classes increased from 60% to 83%, **E**) A normalised confusion matrix displaying the proportions of correct mosquito age class predictions achieved on the held-out Ifakara test set during model training, **F**) Proportions of correct mosquito age class predictions achieved on unseen Glasgow dataset.

When using t-SNE, the pre-trained predicted the age classes in the held-out data (test set) with 83% accuracy but failed to achieve generalisability for the Glasgow data when either 2% or 5% transfer learning was applied, achieving only 50% and 55% accuracy, respectively ((Table 2.1), Fig. 2.6D-F).

Transfer learning also reduced training time while improving the performance of both DL and standard machine learning models in predicting samples from the target population. Transfer learning took less than two minutes for both models to produce the desired results (Table 2.1).

Table 2.1: The performance of deep learning and standard machine learning models for predicting mosquito age classes from the same or alternate insectaries, with and without dimensionality reduction (DR) and transfer learning

Models	Dimensionality reduction (DR) technique	Training data sources	Transfer learning	Base Model runtime	Transfer learning runtime	Predictions for age of mosquitoes from same insectary (Ifakara) - Test accuracy (%)	Predictions for age of mosquitoes from alternate insectary (Glasgow) - unseen data accuracy (%)
CNN-1	No DR	Ifakara	No TL	7.2 seconds/iteration	N/A	99	46
CNN-2	No DR	Ifakara	2% (33 of 1635)	7.2 seconds/ iteration	1 minute	99	100
CNN-3	No DR	Ifakara	5% (82 of 1635)	7.8 seconds/ iteration	2 minutes	99	96
MLP-1	PCA	Ifakara	No TL	6.5 seconds/ iteration	N/A	91	58
MLP-2	t-SNE	Ifakara	No TL	1 seconds/ iteration	N/A	84	58
MLP-3	PCA	Ifakara	2% (33 of 1635)	0.8 seconds/iteration	35 seconds	91	97
MLP-4	PCA	Ifakara	5% (82 of 1635)	0.7 seconds/ iteration	51 seconds	91	96
MLP-5	t-SNE	Ifakara	2% (33 of 1635)	0.7 seconds/ iteration	47 seconds	83	50
MLP-6	t-SNE	Ifakara	5% (82 of 1635)	0.7 seconds/ iteration	49 seconds	83	55
XGB-1	No DR	Ifakara	No TL	645 seconds/iteration	N/A	92	48
XGB-2	No DR	Ifakara	2% (33 of 1635)	975 seconds/iteration	1 seconds	92	98
XGB-3	No DR	Ifakara	5% (82 of 1635)	861 seconds/iteration	1 seconds	92	98
XGB-4	PCA	Ifakara	No TL	60 seconds/iteration	N/A	90	48
XGB-5	t-SNE	Ifakara	No TL	66 seconds/iteration	N/A	68	55
XGB-6	PCA	Ifakara	2% (33 of 1635)	54 seconds/iteration	1 seconds	90	98
XGB-7	PCA	Ifakara	5% (82 of 1635)	54 seconds/iteration	2 seconds	90	97
XGB-8	t-SNE	Ifakara	2% (33 of 1635)	60 seconds/iteration	1 seconds	81	43
XGB-9	t-SNE	Ifakara	5% (33 of 1635)	60 seconds/iteration	1 seconds	82	49

***** CNN–1 to 3: Different versions of convolutional neural network, MLP–1 to 6: Different versions of Multi-Layer Perceptron, XGB-1 to 9: Different versions of XGBoost classifier (standard machine learning), No DR: No dimensionality reduction, PCA: Principal component analysis, t-SNE: t-distributed stochastic neighbour embedding, No TL: No Transfer learning, N/A: Not applicable.

2.4.3 Comparison between deep learning and standard machine learning models in achieving generalisability

The XGBoost classifier (Fig. 2.7A), when trained with Ifakara data only, failed to predict age classes of mosquitoes from the Glasgow insectary, with or without dimensionality reduction (Table 2.1). However, when the classifier was updated with 2% target population samples, the model correctly classified individual mosquito age classes with 98% for both 1–9 days old and 10–17 days old mosquitoes (Fig. 2.7B). Increasing the samples for transfer learning to 5% of the training set had no effect on the accuracies (Table 2.1). However, when t-SNE was used for dimensionality reduction, transfer learning with either 2% or 5% Glasgow samples did not improve the generalisability of the XGBoost classifier (Table 2.1).

(Table 2.2) shows how the performance of deep learning and standard machine learning was evaluated using other metrics such as precision, recall, and F1-scores. When it comes to mosquito age classification, the XGBoost classifier matches the deep learning model in both specificity (precision) and sensitivity (recall).

Table 2.2: Precision, recall, and F1-score of the best deep learning model for classifying mosquito age classes from alternate sources compared to the best standard machine learning algorithm (i.e. XGBoost classifier)

Model name	Age class (Days)	Precision	Recall	F1-score	No. of samples per age class
MLP-3	1-9	0.98	0.97	0.98	895
	10-17	0.97	0.97	0.97	707
XGB-6	1-9	0.98	0.99	0.98	895
	10-17	0.98	0.98	0.98	707

***** MLP–3: Multi-Layer Perceptron trained with PCA as a dimensionality reduction technique and 2% transfer learning, XGB-6: XGBoost classifier trained with PCA as a dimensionality reduction technique and 2% target population samples used for transfer learning.

Further to that, standard machine learning models were trained with 10 iterations, and still the computing runtime were generally shorter than those for CNN models when PCA and t-SNE were used to transform the input data, in some cases by up to 5 times (Table 2.1).

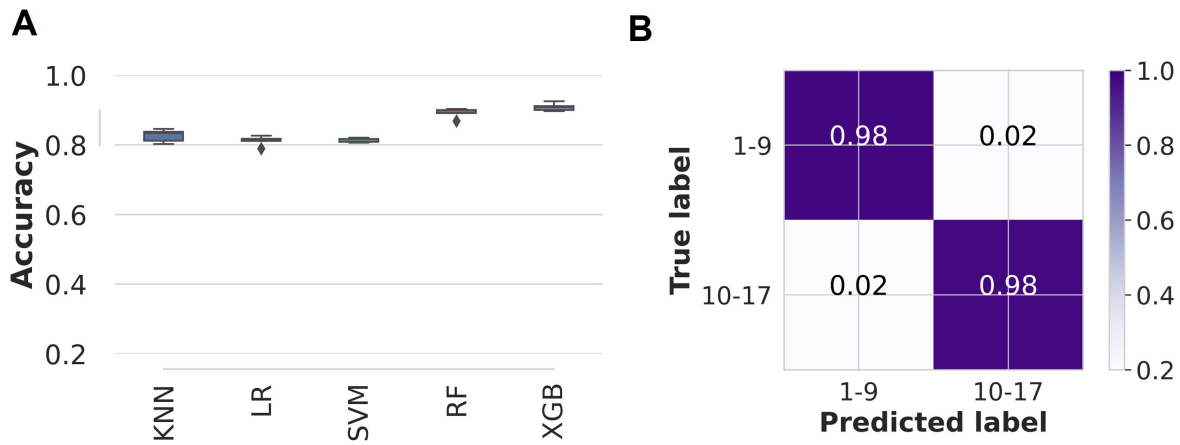

Figure 2.7: Standard machine learning models' predictive accuracies and generalisability when trained with PCA-transformed Ifakara data plus 2% new target population. **A**) Comparison of standard machine learning models for mosquito age classification; KNN: K-nearest neighbours, LR: Logistic regression, SVM: Support vector machine classifier, RF: Random Forest classifier, and XGB: XGBoost. **B**) proportions of correct mosquito age class predictions achieved on unseen Glasgow dataset.

2.5 Discussion

This study demonstrates that transfer learning approaches can substantially improve the generalisability of both deep learning and standard machine learning in predicting the age class of mosquitoes reared in two different insectaries. We evaluated 1635 mosquito spectra from Glasgow-reared mosquitoes and show that using transfer learning and dimensionality reduction techniques could improve machine learning models to predict mosquito age classes from alternate insectaries. Furthermore, reducing the dimensionality of the spectral data reduced computational costs (i.e. computing time) when training the machine learning models. The current study adds to the growing evidence of the utility of infrared spectroscopy and machine learning in estimating mosquito age and survival [60, 90, 94, 147]. In the past, most applications of infrared spectroscopy in estimating mosquito vector survival relied on near-infrared frequencies (12,500 cm^{-1} to 4,000 cm^{-1}). A recent study used mid-infrared spectra (from 4,000 cm^{-1} to 400 cm^{-1} frequencies) and standard machine learning to distinguish mosquito species with up to 82% accuracy, but found lower age prediction accuracy in several alternate settings [59]. González *et al.*, suggested that machine learning under-prediction may be explained by the small training dataset and ecological variability between the training and validation sets [59].

In our study, despite categorising mosquito chronological age into two classes (young: 1-9 day olds and old: 10-17 day olds), deep learning and standard machine learning approaches both remained unable to generalise, even after reducing the dimensionality of

the spectra data. This result is consistent with Siria *et al.*, [61], where CNN under performed as a result of the difference in data distribution between the training and evaluation data driven by non-genetic factors such as ecological variation. When near-infrared spectroscopy was used to predict the age of Anopheles mosquitoes reared from wild populations, a similar limitation was reported [60, 94].

Nonetheless, Siria *et al.*, [61] also observed that using transfer learning to correct the difference data distribution between training and evaluation data improved deep learning generalisation, achieving 94% accuracy in predicting both species and mosquito age classes. Furthermore, in the latter study, the performance of the classifier was improved by incorporating a subset (n = 1,200~1,300 spectra) of the evaluation data into the training data.

The present study shows performing transfer learning using 2% of the spectra from the target domain (33 of 1,635) as well as dimensionality reduction resulted in the improved generalisability of both deep learning and standard machine learning models achieving overall accuracy of 98%. In this case, we expected that all models to which transfer learning was applied would outperform the baseline models as previously demonstrated [61, 148]. However, as the proportion of data from the target domain in the training increased, the performance slightly dropped for the deep learning. The reason for the deterioration in performance after turning the pre-trained\base model with 5% transfer learning could be that the model over-fitted random noise during training, which negatively impacted the performance of these models on unseen data. These results were consistent across multiple random train/test splits. Other studies have proposed alternative transfer learning approaches, such as adaptive regularisation to address cross-domain (i.e., source domain and target domain) learning problems [149], transferring knowledge gained in the source domain during training to the target domain [150], and integrating dimensionality reduction to transform features of the source to ensure data distribution in different domains is minimised [151], such as transfer learning with multi-target regression approach to exploit orthologous genes to capture similarities in metabolic responses in mice and humans [152, 153].

Furthermore, dimensionality reduction was used in conjunction with transfer learning to reduce noise, redundant features, and computational time. Based on our findings, dimensionality reduction alone cannot achieve generalisability of machine learning models. The PCA improved model stability because the eigenvectors of the correlation matrix in PCA provide new axes of variation to project new data while preserving the original distance between the points in the data. The model with t-SNE as a dimensionality reduction technique failed to achieve generalisability on the new data, the reason for poor performance could be t-SNE is a probabilistic technique with a non-convex cost function [140], causing the output to differ from multiple runs, and may not preserve the original distances

between the points in the data. In this study, PCA is considered a better choice than other dimensionality reduction technique for training machine learning models from spectra data because it is simple to implement, computationally efficient, and produces good results.

Furthermore, incorporating dimensionality reduction substantially reduces model training time and thus, computational requirements. When compared to models trained without dimensionality reduction, the computing runtime for models trained with dimensionality reduction were less than five-fold. Moreover, transfer learning in general was fast, tuning the pre-trained models in under two minutes on our machine (standard laptop). This makes the technique applicable and reproducible even to users with low computing power and capacity providing they have access to pre-trained models.

This study only included *An. arabiensis* reared in the laboratory from two insectaries. Future research should put the techniques to the test with samples from more laboratories, field settings, and mosquito species, as these factors can affect the model's predictive capacity. The optimal ratio of transfer learning data required to achieve best generalisability in predicting mosquito age class has yet to be determined, so future studies could investigate this gap. Furthermore, because dimensionality reduction reduced the computational requirements in this study, we suggest that clustering spectra with algorithms such as PCA can be a beneficial strategy for models trained on MIRS.

2.6 Conclusion

This study found that using transfer learning and dimensionality reduction with principal component analysis (PCA) improved the generalisability of machine learning models for predicting mosquito age classes from 56% to ≥97%. This suggests that these techniques could be scaled up and further evaluated to determine the age of mosquitoes from different populations. In addition, when dimensionality reduction and transfer learning are used, simpler machine learning algorithms, such as the XGBoost classifier, can reduce computational time while still achieving performance close or equal to deep learning. This could help entomologists reduce the amount of time and work required to dissect large numbers of mosquitoes. Overall, these approaches have the potential to improve model-based surveillance programs, such as assessing the impact of malaria vector control tools, by monitoring the age structures of local vector populations.

For future research, our goal is to create a large database of spectra data and use transfer learning to build a pipeline that can predict the age of wild malaria mosquitoes across different populations in order to support vector surveillance in malaria-endemic areas. Here we have presented a new technique that uses transfer learning and dimensionality reduction to improve the generalisability of machine learning predictions. However, the

optimal proportion of new data from target populations required for generalisability is still unknown, and warrants further optimisation.

2.7 Ethical approval

All methods in this study were performed in accordance with the relevant guidelines and regulations from IHI, National Institutes of Medical Research (NIMR), and UofG. At IHI, ethical approval for the study was obtained from the IHI Institutional Review Board (Ref. IHI/IRB/EXT/No: 005-2018), and NIMR Ref: NIMR/HQ/R.8c/Vol.II/880. At the UofG, Ethical approval for the supply and use of human blood for mosquito feeding was obtained from the Scottish National Blood Transfusion Service committee for governance of blood and tissue samples for non-therapeutic use (submission Reference No 1815).

2.8 Availability of data and materials

The mid-infrared spectral data generated and/or analysed during the current study are deposited and available in the https://doi.org/10.5525/gla.researchdata.1235.

2.9 Competing interests

The authors declare that they have no competing interests.

2.10 Funding

This research was supported by the Medical Research Council (MRC) grant (Grant No. MR/P025501/1). EPM and DJS were also supported by the Wellcome Trust International Masters Fellowships in Tropical Medicine and Hygiene, Grant Nos. WT214643/Z/18/Z and WT 214644/Z/18/Z respectively.

2.11 Authors contributions

EPM, SAB, DJS, FB, MGJ and FOO designed the study. DJS supported in data collection semi-field experiments. EPM performed data analysis. JM provided technical support to EPM during data analysis. EPM wrote and revised the manuscript. EPM, SAB, IHM, PS, FOO, and FB reviewed and revised the manuscript. All authors read and approved the final manuscript.

Chapter 3

Rapid Classification of Epidemiologically Relevant Age Categories of the Malaria Vector, *Anopheles funestus*

Emmanuel P. Mwanga[1,2*], Doreen J. Siria[1,2], Issa H. Mshani[1,2], Sophia H. Mwinyi[1,2], Said Abbas[1], Mario Gonzalez Jimenez[2,3], Klaas Wynne[3], Francesco Baldinit[1,2], Simon A. Babayan[2], Fredros O. Okumu[1,2,4,5].

1. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, P.O. Box 53, Morogoro, Tanzania.
2. School of Biodiversity, One Health and veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, UK.
3. School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK.
4. School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
5. School of Life Science and Bioengineering, The Nelson Mandela African Institution of Science and Technology, P. O. Box 447, Arusha, Tanzania.

DOI: https://doi.org/10.1186/s13071-024-06209-5

3.1 Abstract

Background: Accurately determining the age and survival probabilities of adult mosquitoes is crucial for understanding parasite transmission, evaluating the effectiveness of control interventions and assessing disease risk in communities. This study was aimed to demonstrating rapid identification of epidemiologically relevant age categories of *Anopheles funestus*, a major Afro-tropical malaria vector, through the innovative combination of infrared spectroscopy and machine learning, instead of the cumbersome practice of dissecting mosquito ovaries to estimate age based on parity status.

Methods: *An. funestus* larvae were collected in rural south-Eastern Tanzania and reared in the insectary. Emerging adult females were sorted by age (1-16 day-olds) and preserved

using silica gel. PCR confirmation was conducted using DNA extracted from mosquito legs to verify the presence of *An. funestus* and eliminate undesired mosquitoes. Mid-infrared spectra were obtained by scanning the heads and thoraces of the mosquitoes using an ATR FT-IR spectrometer. The spectra (N = 2,084) were divided into two epidemiologically relevant age groups: 1-9 days (young, non-infectious) and 10-16 days (old, potentially infectious). The dimensionality of the spectra was reduced using principal component analysis, then a set of machine learning and multi-layer perceptron (MLP) models were trained using the spectra to predict the mosquito age categories.

Results: The best performing model, XGBoost, achieved an overall accuracy of 87%, with classification accuracies of 89% for young and 84% for old *An. funestus*. When the most important spectral features influencing the model performance were selected to train a new model, the overall accuracy increased slightly to 89%. The MLP model, utilising the significant spectral features, achieved higher classification accuracies of 95% and 94% for the young and old *An. funestus*, respectively. After dimensionality reduction, the MLP achieved 93% accuracy for both age categories.

Conclusion: This study shows how machine learning can quickly classify epidemiologically relevant age groups of *An. funestus* based on their mid-infrared spectra. Having been previously applied to *An. gambiae*, *An. arabiensis* and *An. coluzzii*, this demonstration on *An. funestus* underscore the potential of this low-cost, reagent-free technique for widespread use on all the major Afro-tropical malaria vectors. Future research should demonstrate how such machine-derived age classifications in field collected mosquitoes correlate with malaria in human populations.

Key terms: Malaria, *Anopheles funestus*, deep learning, machine learning, Ifakara health institute, mid-infrared Spectroscopy

3.2 Background

Despite significant investments in malaria control and research, there were still an estimated 249 million malaria cases and 619,000 deaths in 2021 globally, a significant majority of which occurred in sub-Saharan Africa [6]. Other than the poor economic conditions and weak health systems, the continued high burden of malaria in Africa is attributable to key biological threats, notably malaria parasite resistance to drugs [154–156], vector resistance to insecticides [157, 158], increasing occurrence of malaria parasites evading detection by rapid diagnostic tests [159–163], and disruptions from major disease outbreaks such as Ebola and COVID-19 [2, 164, 165]. Effective vector control, primarily with insecticide treated nets (ITNs) and indoor residual spraying (IRS), has been the most important component of malaria control in Africa [12]. However, its continued effectiveness requires

active innovation to address the current threats, and improved understanding of the major vector species in different settings.

Anopheles funestus is one of the four main malaria vector species in sub-Saharan Africa, the others being *An. gambiae*, *An. arabiensis* and *An. coluzzii*, and also one of the most widespread [127–129, 133]. *An. funestus* is particularly important in East and Southern Africa, where it is becoming the dominant malaria vector. For example, in parts of Tanzania, *An. funestus* is reported to be responsible for 86-97% of all new malaria infections [128, 130, 131, 134]. Its dominance is due to multiple factors, including i) being highly anthropophilic, and thus preferring to bite humans over other vertebrates [128, 136], ii) being highly endophilic, i.e. preferring to bite inside human dwellings than outside [166], iii) having significantly higher survival rates than other species [167], iv) being resistant to commonly used insecticides [127, 128, 168] and v) preferentially breeding in perennial habitats with year-round productivity [23]. Given its importance and dominance in malaria transmission systems, vector surveillance programs in the respective countries should be designed with special attention to this vector species.

Besides evaluating biting densities and *Plasmodium* infection rates, accurately determining the age and survival of *An. funestus* is crucial for monitoring transmission dynamics and assessing the effectiveness of vector control interventions such as ITNs and IRS. Dissection of mosquito ovaries is still the main entomological technique for estimating the age of vector populations [53]. The dissections are usually performed under light microscopes to assess the reproductive history, specifically the parity status, of the mosquitoes. This involves observing whether the ovaries contain coiled tracheolar skeins (indicating non-parous mosquitoes) or stretched-out tracheoles (indicating parous mosquitoes). Non-parous mosquitos are considered young in this case, whereas parous mosquitos are considered old and may carry the malaria parasites, having had multiple blood-feedings [53]. Unfortunately, these dissections tend to be laborious and time-consuming, especially when dissecting large numbers of mosquitoes, and are impractical on a large scale.

Furthermore, the reliability of mosquito dissections is limited by their reproductive history. For instance, a female mosquito can have more than one blood meal but still not oviposit, a scenario known as gonotrophic discordance or pre-gravid blood-meal [55]. Moreover, since the gonotrophic cycles of *Anopheles* mosquitoes can be as short as 2–3 days under optimal climatic conditions [169, 170], it is possible for parous mosquitoes to be relatively young, and in rare cases, nulliparous mosquitoes to be several days old due to the scarcity of blood meals (e.g. when ITNs coverage and usage is high). Therefore, using parity alone to distinguish between epidemiologically distinct age categories of adult mosquitoes, especially in the context of malaria transmission, which requires 10-14 days of incubation [20], is not always realistic.

All these concerns suggest the need for alternative age-grading techniques that are easy to perform cheaply at scale and can provide accurate representations of epidemiologically important mosquito age categories and populations. The alternative mosquito age-grading methods currently include the analysis of cuticular hydrocarbon patterns in a gas chromatograph [171] and gene transcription [72, 172, 173]. Near-infrared spectroscopy (NIRS) (12,500 cm^{-1} to 4,000 cm^{-1} frequencies) [174], which involves passing infrared light through a mosquito sample to measure absorbance or reflectance of the organic compound functional groups, has also been used to estimate ages for various mosquito species of both laboratory-reared and wild collected mosquitoes [60, 94, 95, 147, 175–177].

More recently, mid-infrared spectroscopy (MIRS) has been used to predict and estimate mosquito age, recording the biochemical composition of mosquito samples at longer wavelength frequencies [59, 61, 178]. In addition, machine learning (ML) techniques, including convolutional neural networks, have been utilised to differentiate MIRS spectra associated with distinct mosquito ages and species in both laboratory and wild mosquitoes [61, 178]. The infrared based systems have so far been successful for various applications on three of the four main African malaria vectors (i.e. *An. gambiae s.s*, *An. arabiensis* and *An. coluzzii* [61], but have yet to be demonstrated for *An. funestus*. The goal of this study was therefore to test whether a similar ML-MIRS approach could classify adult female *An. funestus* mosquitoes derived from wild-caught larvae into two epidemiologically relevant age categories: young (0-9 days old, too young to have mature *Plasmodium* sporozoites in their salivary glands) and old (10 days or older, potentially carrying mature *Plasmodium* sporozoites given the right climatic conditions), factoring in a parasite incubation period of 10-14 days.

3.3 Methods

3.3.1 Mosquito collection

Third and fourth instar mosquito larvae were collected from known aquatic habitats of *An. funestus* in five different villages in Southeastern Tanzania, namely Tulizamoyo (8.3669°S, 36.7336°E), Kilisa (8.3721°S, 36.5584°E), Lupiro (8.38333°S, 36.66667°E), Ikwambi (7.9833°S, 36.8184°E), and Ruaha (8.9068°S, 36.7185°E). The larvae were transported to the vector biology laboratory (VectorSphere), at Ifakara Health Institute for further rearing. The larvae were kept in water from their natural breeding habitats and were fed Tetramin® fish food. Once they pupated, the pupae were separated from the larvae and placed in emergence cages. The emergent adult mosquitoes were maintained at 26-28°C, 70-85% relative humidity and a 12:12 hour light/dark photoperiod, on a 10% sugar solution diet.

3.3.2 Mosquito preservation and scanning

The female adults were collected and individually preserved according to their age, from 1 to 16 days old. A total of 2084 mosquitoes were collected. The female mosquitoes were killed using chloroform and subsequently stored in separate 1.5 ml microcentrifuge tubes containing silica gel for desiccation. The heads and thoraces of the individual female mosquitoes were scanned using an Attenuated total reflection - Fourier transform infrared spectrometer (ATR – FT-IR) to obtain mid-infrared spectra with a resolution of 2 cm^{-1} at 4,000 - 400 cm^{-1} frequencies as previously described, complete with background spectral calibration [59, 114, 115]. For each sample, 16 sample scans were averaged to obtain the primary output spectrum [61].

3.3.3 Mosquito identification

Though the field collections had been done in known *An. funestus* habitats, it was necessary to confirm the identity of the mosquitoes and eliminate any unwanted species. This was accomplished primarily by morphology-based taxonomy using keys of Afro-tropical *Anopheles* [179] but was complemented by PCR identification to sort between sibling species in the *An. funestus* group. Wild *An. funestus* complex DNA was extracted from the two legs of adult female mosquitoes. The two legs of an individual *An. funestus* mosquito were placed separately in a 1.5 ml micro-centrifuge tube, followed by 20μL of TE buffer (Tris -EDTA), and incubated at 95°C for 15 minutes. PCR was then used to differentiate *An. funestus* from other sibling species, using species-specific primers targeting the non-coding region of ITS2 using the protocol by Koekmoer *et al.*, [180]. The PCR reaction was performed in a 25μL volume, consisting of a PCR mixture of 2.5μL 10x reaction buffer, 25mM $MgCl_2$, 10pmol/μL of each primer, 8mM of each dNTP, 5 units of thermo-stable Taq DNA polymerase, and 3μL of DNA template. The PCR products were analysed by electrophoresis in 2.5% agarose gel stained with classic view DNA dye for visualisation of DNA bands. Only *An. funestus* mosquitoes were considered for further analysis, and any other species discarded.

3.3.4 Machine learning

Mosquito spectra with low intensity, abnormal background or atmospheric interferences (with water vapor and carbon dioxide) were discarded [59]. The data from the remaining spectra (N = 2,084) were processed and analysed in Python 3.9, using Scikit-learn [141],

and Tensorflow 2.0 [144,145]. The data were rescaled using the Standard Scaler algorithm, with a mean of 0 and a standard deviation of 1.

Using the algorithm stratified shuffle split, the dataset was split into training (n = 1,875) and test/unseen (n = 209) sets. To train the supervised ML models, *An. funestus* ages were used as training labels. *An. funestus*, ranging from 1 to 16 days old, were divided into two epidemiologically relevant age categories taking into consideration the incubation period of malaria parasites of 10-14 days [20]. The first group included *An. funestus* that were between 1 to 9 days old and were considered young and incapable of transmitting malaria (i.e., non-infectious age group). The second group included *An. funestus* that were between 10 to 16 days old and were considered old enough to be capable of transmitting malaria given the right environmental conditions (i.e., potentially infectious).

Multiple standard machine learning (ML) classifiers, including K-nearest neighbours (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost), were compared to determine which model predicted the data with the highest classification accuracy. The best-performing model was further optimised by fine-tuning its hyper-parameters. The top 100 spectral features (wavenumbers) with the most influence on the model predictions were identified and utilised to reduce the dimensionality of the spectra data, followed by retraining of the best ML classifier.

Moreover, two Multi-layer Perceptron (MLP) models were trained by reducing the dimensionality of the spectra data using different inputs: 1) the top 100 features extracted from the best performing ML classifier, and 2) principal components using Scikit-learn library. Both MLP models had six fully connected layers, each containing 500 neurons, to enable the model to learn from the network's weights as previously demonstrated [178]. To prevent overfitting, a dropout layer with a rate of 0.5 was used, and early stopping was implemented when the validation loss could no longer improve after 400 iterations [181,182]. The model's performance was evaluated using K-fold cross-validation ($k = 5$) to ensure an unbiased assessment of the standard ML and MLP models, as previously described [178].

To assess the ability of the optimised models to identify all positive instances and avoid false negatives, the recall score (i.e. sensitivity or true positive rate) was estimated as a ratio of correctly age-classified *An. funestus* to the total number of *An. funestus* in the respective age category in the dataset. Moreover, to measure the ability of the models to avoid false positives, the precision score (i.e. the positive predictive value) was estimated as a ratio of correctly age-classified *An. funestus* to the total number of predicted positive instances of the respective age categories. Lastly, we calculated the F1-score, which balances both precision and recall scores by giving equal weight to both measures. This score provides a single value that represents the overall performance of the model in terms of its ability to

correctly classify positive and negative cases. A higher F1-score signifies a better model performance, where a maximum value of 1 represents flawless precision and recall.

3.4 Results

3.4.1 Predicting *An. funestus* age classes using standard machine learning models

In the initial comparison of standard ML models, XGBoost emerged as the best classifier with the highest prediction accuracy and lowest standard deviation, achieving 84% accuracy (Fig. 3.1A). After optimising the parameters, the XGBoost model was able to classify spectra that were previously unseen with an overall accuracy of 87%. It achieved accuracies of 89% and 84% for young (1-9 days old) and old (10-16 days old) *An. funestus* females respectively (Fig. 3.1B). The recall scores (i.e. sensitivity or true positive rates) of this model were 0.89 and 0.84 for the young and old mosquitoes respectively, while its precision scores (i.e. the positive predictive value) were 0.87 for both age categories (Table 3.1).

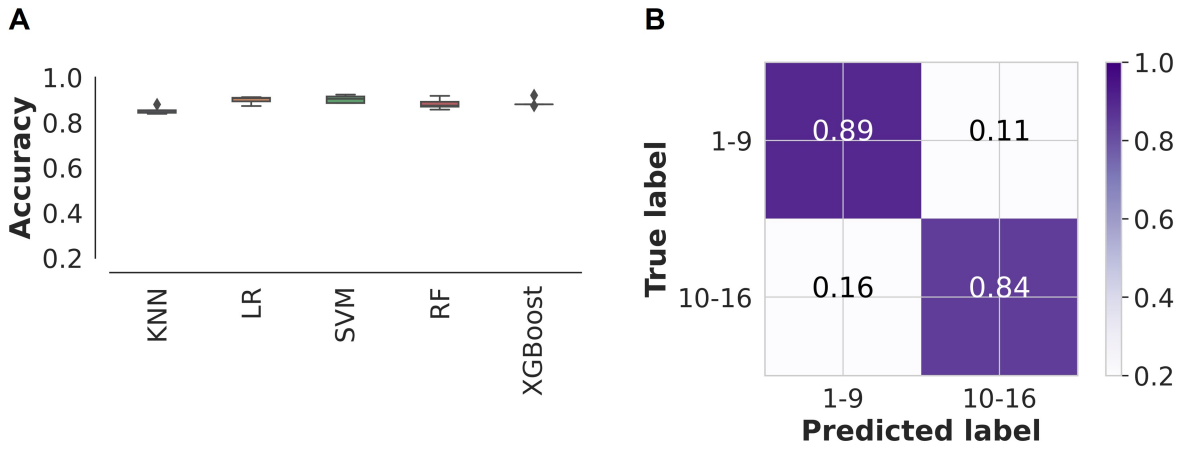

Figure 3.1: Machine learning prediction of *An. funestus* age classes. **A**) Comparison of standard ML classifiers in predicting *An. funestus* age classes; KNN: K-nearest neighbours, LR: Logistic regression, SVM: Support vector machine, RF: Random Forest, XGBoost: Gradient boosting, MLP: Multilayer perceptron. **B**) Confusion matrix for predicting the age class of *An. funestus* using XGBoost on an unseen dataset, results for the ML trained with all spectra features.

From the initial XGBoost model, we identified the spectral features that were most important for the prediction. This analysis aimed to reduce the number of training features and enhance the accuracy of the model during retraining (Fig. 3.2A). When the XGBoost classifier was retrained with the top 100 features, the classification accuracy increased to

89%, correctly predicting young and old *An. funestus* females with 92% and 85% accuracies respectively (Fig. 3.2B).

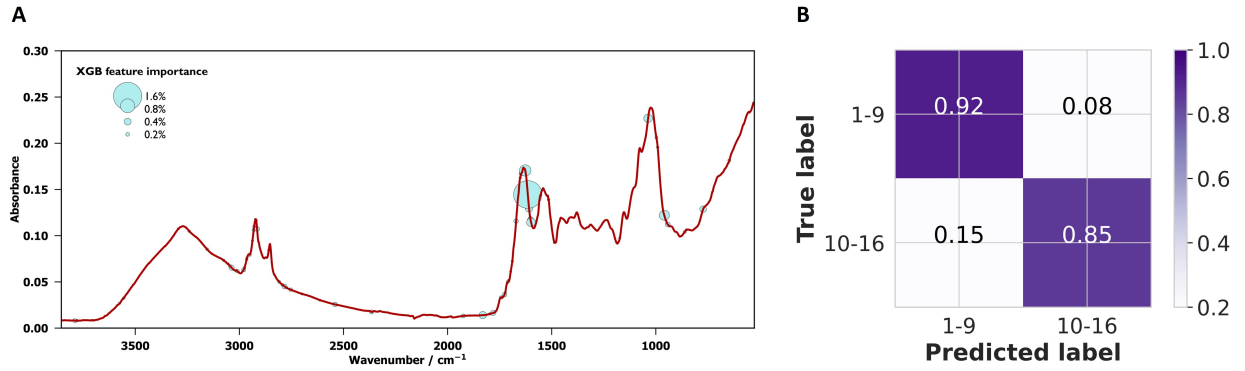

Figure 3.2: **A**) Relative importance of XGBoost features that have the most influence in predicting the age classes of *An. funestus*. **B**) Confusion matrix for predicting the age class of *An. funestus* using XGBoost on an unseen dataset, the results for the ML retrained with important features/wavenumbers ($n = 100$) identified by the initial XGBoost model.

3.4.2 Prediction of *An. funestus* age classes using Multi-layer perceptron (MLP) models

We explored the possibility of improving the accuracy by training the MLP classifier using the important wavenumbers ($n = 100$) identified in the XGBoost predictions. As a result, the MLP achieved an improved accuracy of 94.5% in the unseen test data (Fig. 3.3A), correctly distinguishing between young and old *An. funestus* females with accuracies of 95% and 94%, respectively (Fig. 3.3B).

Lastly, in a previous study, we presented evidence that employing PCA with eight components effectively reduces the dimensionality of the spectra data [178]. This reduction in dimensionality not only preserved a substantial portion of the data variability, but also mitigated overfitting while enhancing the signal-to-noise ratio. By utilising a reduced set of features, we trained the MLP model to improve its predictive performance [178]. In the present study, when PCA was utilized to reduce the dimensionality of the spectra data, the MLP classifier achieved an overall accuracy of 93% for both young and old *An. funestus* mosquitoes (Fig. 3.3C).

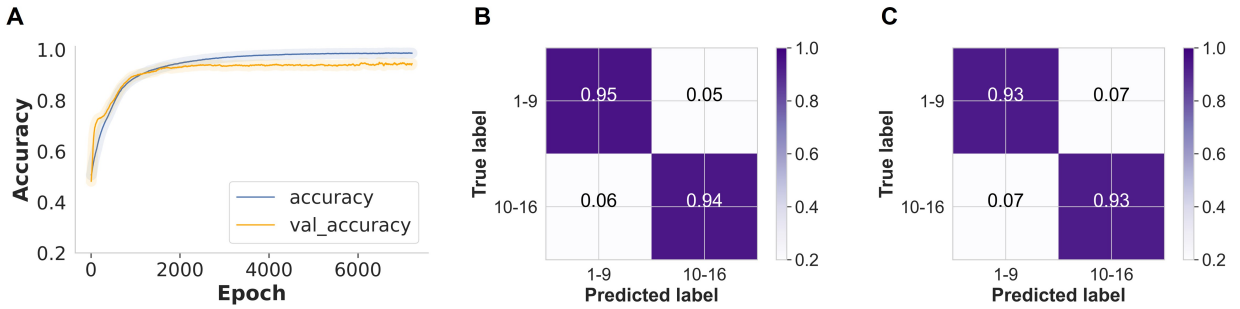

Figure 3.3: **A**) MLP Training and validation accuracy for *An. funestus* age classes as training time increases (epoch; number of iterations over the entire dataset during the training process, i.e. seconds/iterations). Confusion matrix for predicting the age class of *An. funestus*; Panel **B** shows the results for the MLP trained with important features/wavenumbers ($n = 100$) identified by the XGBoost. Panel C shows the results for the MLP method trained with eight principal components.

Table 3.1: Precision, recall, and F1-score of XGBoost and multi-layer perceptron (MLP) models for predicting age categories of *An. funestus*

Model	Age classes	Precision	Recall	F1-score	No. test samples
XGBoost 1	1 - 9	0.87	0.89	0.88	113
	10 - 16	0.87	0.84	0.86	96
XGBoost 2	1 - 9	0.88	0.92	0.90	113
	10 - 16	0.90	0.85	0.88	96
MLP 1	1 - 9	0.95	0.95	0.95	113
	10 - 16	0.94	0.94	0.94	96
MLP 2	1 - 9	0.94	0.93	0.93	113
	10 - 16	0.92	0.93	0.92	96

***** XGBoost 1: Trained with all MIRS wavenumbers (n = 1,665), XGBoost 2: Trained with spectral features extracted based on feature importance summaries (n = 100), MLP 1: Trained with spectral features extracted based on feature importance summaries (n = 100), MLP 2: Trained with PCA as a dimensionality reduction technique.

3.5 Discussion

An. funestus mosquitoes are currently the major vector of malaria transmission in Tanzania, accounting for over 80% of malaria transmission [128, 130, 131, 134]. *An. funestus* tends to have better survival rates [167], and is generally a slow-growing mosquito, which adds to the challenge of studying its demographic characteristics and how these might influence pathogen transmission. Here, we present a rapid age-grading technique that

has potential to replace the traditional methods like ovarian dissections, which are time-consuming and challenging to apply on a large scale. Using 2084 spectra data, we trained machine learning models that classify the epidemiologically relevant age groups of *An. funestus* mosquitoes reared from wild larvae using water from the same habitats, but under laboratory conditions. The models correctly distinguished between the young *An. funestus* females (1-9 days old) and the older ones (10-16 days old) based on the MIR spectra indicative of the varying biochemical composition of the mosquito cuticles [116]. While this was the first demonstration of the effectiveness of this technique for predicting the age of *An. funestus* mosquitoes, the approach of combining infrared spectra and machine learning models has been widely demonstrated for predicting different indicators including age, blood meals, infection status, and insecticide resistance profiles of other *Anopheles* species [61, 114, 115]. If validated on field collected adults, these findings could be a step towards wider applications of this approach for malaria vector surveillance in settings with different vector species.

In settings such as rural south-eastern Tanzania where *An. funestus* is the dominant malaria vector [128, 130] it is particularly important that vector surveillance programs are expanded to include this vector species. Indeed, the successful demonstration that this technique on *An. funestus*, which is one of the most efficient and also most widespread malaria vectors in Africa [183], expands the range of utility of this technique for a much broader application for malaria vector surveys in different parts of Africa.

One of the key concerns regarding previous applications of MIRS-ML based approaches for entomological assessments is that, with exception of some cases [61], these methods have been rarely validated for wild-caught malaria vectors in field settings. Here, *An. funestus* mosquitoes were collected as larvae from various villages and breeding habitats, to account for genetic variation, variation in larval food sources and microbiome, and to maintain some characteristics of the natural ecosystems. The success of this analysis and the high accuracies obtained may therefore be indicative of the potential of the approach for predicting key mosquito attributes in field settings. However, it is unknown whether specific climatic factors could influence the prediction and generalisability of MIRS-ML approach. Future studies should therefore test the generalisability of this approach across different populations of wild mosquitoes.

This study classified mosquitoes only as young (1-9 days old) or old (10-16 days old) and did not attempt to classify them at specific chronological ages because the sample size was not large enough to test it. However, the chosen age classes represent the typical epidemiological distinction relevant to the transmission of malaria parasites, which, under standard climatic conditions, requires that a vector must be at least 10 days old [20]. However, it may fail to capture variations in MIR spectra or the small biochemical changes that occur within a mosquito cuticle after each ageing day (such as chronological age from

1 up to 16) [59]. Moreover, it has been demonstrated that calibrating machine learning models based on physiological age (which considers key life cycle processes such as blood-feeding and oviposition) may be more useful than simply relying on chronological age classifications [60, 90]. In our study, mosquitoes were all sugar fed, and therefore physiological age was not assessed. Future efforts should assess key differences in these approaches and evaluate models trained on biological age and chronological age to determine which ones are most practical and most generalisable. An obvious next step is therefore to investigate any correlations that might exist between the machine-classified age categories and the epidemiology of malaria in human populations.

To improve the classification accuracy of our model, the XGBoost feature importance was relied upon to reduce the number of spectral features from 1,665 to 100. This dimensionality reduction significantly lowered the noise and redundant features in the MIR spectra data. The important features were mostly associated with proteins, with the most influential peak (1,700 cm^{-1}) being the band associated with the amide bond from proteins. The region around 3,000 cm^{-1}, which is also related to proteins, was also found to be important in the model prediction. This implies that the model is learning from protein-based biological traits that vary depending on the age of the mosquito [61]. Moreover, when PCA was used to reduce the dimensionality of the spectra from 1,665 features to eight principal components [178], the prediction accuracy matched that of the MLP model trained with the top 100 biological features as identified from the XGBoost model. This suggests that machine learning models may perform better when trained with fewer features that explain more variation in the data, rather than many redundant features that introduce noise into the model. Moreover, as observed previously, reducing the dimensionality of the spectra data reduces the computational resources needed to train machine learning models [178].

Future research should investigate the effects of rearing wild *An. funestus* larvae in the insectary on the predictive accuracies of MIRS-ML approach for mosquito age-classification as this could impact the generalisability of the findings.

3.6 Conclusion

This study demonstrates the classification of adult female *An. funestus* into distinct and epidemiologically relevant age categories using a MIRS-ML approach. In conjunction with prior research conducted on other *Anopheles* mosquitoes, this study suggests that the applicability of this approach can be extended to evaluate various entomological attributes in *An. funestus*. The MIRS-ML approach proves to be quick, cost-effective, and has the potential to significantly enhance *An. funestus* surveillance efforts, thereby contributing valuable insights to national malaria control programs, particularly in resource-constrained settings where this vector is highly prevalent. Nonetheless, further research is needed to

validate the MIRS-ML approach in field conditions, using adult *An. funestus* populations and other vector species within malaria-endemic communities, and to examine how the machine-classified age categories correlate with the epidemiological strata of malaria in human populations.

3.7 Ethics approval and consent to participate

Ethical approval for this study was obtained from Ifakara Health Institute Institutional Review Board (Ref. IHI/IRB/EXT/No: 005-2018), and from the Medical Research Coordinating Committee (MRCC) at the National Institute of Medical Research (NIMR), Ref: NIMR/HQ/R.8c/Vol. II/880.

3.8 Code, data, and materials availability

The mid-infrared spectral datasets generated and analysed during the current study, as well as code for the analyses is available at https://github.com/MwangaEP/Funestus-age.

3.9 Competing interests

The authors declare that they have no competing interests.

3.10 Funding

This study was supported by a Howard Hughes Medical Institute (HHMI)-Gates International Research Scholarship (Grant No. OPP1099295) awarded to FOO and the Medical Research Council (MRC) [MR/P025501/1] awarded to FB. EPM was supported by the Wellcome Trust Masters Fellowship in Tropical Medicine and Hygiene (Grant No. 214643/Z/18/Z). FB is supported by the Academy Medical Sciences Springboard Award (Ref:SBF007/100094). SAB is supported by the Bill and Melinda Gates Foundation (INV-030025) and Royal Society (ICA/R1/191238).

3.11 Authors' contributions

EPM, DJS, SB, FB, and FOO conceived the study. EPM, SA, FOO, and DJS developed the study's protocol. DJS collected the data, EPM carried out data analysis and ML training. EPM wrote the manuscript. EPM, DJS, SHM, IHM, MGJ, KW, SB, FB and FOO reviewed and edited drafts of the manuscript. All authors have read and approved the final manuscript.

Chapter 4

Rapid Assessment of the Blood-feeding Histories of Wild-caught Malaria Mosquitoes Using Mid-infrared Spectroscopy and Machine Learning

Emmanuel P. Mwanga[1,2*], Idrisa S. Mchola[1], Faraja E. Makala[1], Issa H. Mshani[1,2], Doreen J. Siria[1,2], Sophia H. Mwinyi[1,2], Said Abbasi[1], Godian Seleman[1], Jacqueline N. Mgaya[1], Mario González-Jiménez[3], Klaas Wynne[3], Maggy T. Sikulu-Lord[4], Prashanth Selvaraj[5], Fredros O. Okumu[1,2,6,7] , Francesco Baldini[1,2], Simon A. Babayan[2]

1. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, Morogoro, Tanzania.
2. School of Biodiversity, One Health and veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, UK.
3. School of Chemistry, University of Glasgow, Glasgow, G12 8QQ, UK.
4. Faculty of Science, School of the Environment, The University of Queensland, Brisbane, QLD, Australia
5. Institute for Disease Modelling, Bill and Melinda Gates Foundation, Seattle, USA.
6. School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
7. School of Life Science and Bioengineering, The Nelson Mandela African Institution of Science and Technology, P. O. Box 447, Arusha, Tanzania.

DOI: https://doi.org/10.1186/s12936-024-04915-0

4.1 Abstract

Background: The degree to which *Anopheles* mosquitoes prefer biting humans over other vertebrate hosts, i.e. the human blood index (HBI), is a crucial parameter for assessing malaria transmission risk. However, existing techniques for identifying mosquito blood meals are demanding in terms of time and effort, involve costly reagents, and are prone to

inaccuracies due to factors such as cross-reactivity with other antigens or partially digested blood meals in the mosquito gut. This study demonstrates the first field application of mid-infrared spectroscopy and machine learning (MIRS-ML), to rapidly assess the blood-feeding histories of malaria vectors, with direct comparison to PCR assays.

Methods and Results: Female *Anopheles funestus* mosquitoes (N = 1,854) were collected from rural Tanzania and desiccated then scanned with an attenuated total reflectance Fourier-transform Infrared (ATR-FTIR) spectrometer. Blood meals were confirmed by PCR, establishing the 'ground truth' for machine learning algorithms. Logistic regression and multi-layer perceptron classifiers were employed to identify blood meal sources, achieving accuracies of 88% and 90%, respectively, as well as HBI estimates aligning well with the PCR-based standard HBI.

Conclusions: This research provides evidence of MIRS-ML effectiveness in classifying blood meals in wild *Anopheles funestus*, as a potential complementary surveillance tool in settings where conventional molecular techniques are impractical. The cost-effectiveness, simplicity, and scalability of MIRS-ML, along with its generalisability, outweigh minor gaps in HBI estimation. Since this approach has already been demonstrated for measuring other entomological and parasitological indicators of malaria, the validation in this study broadens its range of use cases, positioning it as an integrated system for estimating pathogen transmission risk and evaluating the impact of interventions.

Key terms: *Anopheles*, human blood index, machine learning, transfer learning, VectorSphere

4.2 Background

Effective entomological surveillance requires systematic collection, analysis, and interpretation of data on insects that transmit pathogens in different localities. It is essential for assessing risks and guiding the planning and implementation of vector control strategies, as well for monitoring, and evaluation of those strategies [184]. The likelihood of pathogen transmission can vary widely, depending on factors such as the presence of competent vectors, favourable climatic conditions, the presence of vulnerable human populations and the presence of other vertebrate hosts, which may sustain the vector populations [184]. Other factors may include the diversity of vector species in the area, their population dynamics, their behaviours in and around human dwellings such as the timing and location of biting, their resting behaviours and host preferences of these vectors [136, 185].

Anopheles mosquitoes are considered particularly hazardous due to their propensity to feed on, and thus transmit pathogens to, humans, notably malaria, which causes approximately 620,000 deaths and about 250 million cases annually [6] . Compared to mosquitoes from other regions, the Afro-tropical malaria vectors are particularly dangerous in this regard due to their comparatively greater preference for humans over other vertebrates [136]. This attribute, which is generally estimated as the human blood index, has been considered an important measure of the stability of malaria in different settings [41]; and is known to be highest in major malaria vectors, including *Anopheles gambiae*, *Anopheles funestus* and *Anopheles coluzzii*, which appear to be particularly well adapted synanthropes [27]. Following closely is *Anopheles arabiensis*, which can be an opportunistic vector species capable of blood-feeding readily on either humans or cattle, depending on availability [136, 185, 186]. Consequently, while this behaviour poses a notable risk for the transmission of zoonotic pathogens in addition to malaria, *An. arabiensis* is also a far less competent vector of malaria than either *An. gambiae*, *An. funestus* or *An. coluzzii* [128, 130, 131, 187].

While anthropophagy (i.e., preference for feeding on humans) in malaria vectors can be augmented by the degree of endophily (i.e., preference for indoor resting), this behaviour can also be attenuated under high degrees of exophily (i.e., preference for outdoor resting). For example, *An. funestus* is known for being both highly anthropophilic and highly endophilic [128, 136], enforcing its major role in malaria transmission [128, 130] though there are settings where it is known to bite outdoors early in the morning [188, 189] or to feed on non-human hosts [190]. On the other hand, mosquitoes that rest indoors are more likely to feed on human host, while mosquitoes that prefer to rest outdoors are more likely to feed on non-human host [136, 191]. This might be due to mosquitoes feeding on the first host they encounter when presented with multiple hosts in the same environment [186], or to the use of bed nets preventing access to human hosts [192, 193]. Overall, accurate determination of the blood-feeding histories of malaria vector species is an important indicator of their feeding behaviour, their role in ongoing malaria transmission and the overall risk exposure of people within those settings.

Methods for investigating the blood-meal sources in mosquitoes include several techniques: the precipitin test observes the formation of a white precipitate resulting from the interaction between a saline extract of the blood meal and a suitable antiserum from a known host, indicating the presence of an antigen-antibody interaction [194]; microsphere assays is a molecular-based assay involving uniquely labelled microspheres with host species-specific capture probes to detect host blood meals [195]; microsatellite assays analyse short tandem repeat sequences in the mosquito's DNA to identify blood sources based on unique genetic markers [196]; enzyme-linked immunosorbent assays (ELISA) detect immunoglobulin G (IgG) from blood-fed mosquito samples [68]; and polymerase chain reactions (PCR) target mitochondrial cytochrome b to identify arthropod blood

meal sources [69]. ELISA and PCR, the most common techniques for studying host blood meals in mosquitoes, have played a crucial role in understanding mosquito host preference since the early 1980s and emerged as powerful tools due to their sensitivity [68,69,197–200]. These methods have evolved over time with modification to enhance accuracy and efficiency. ELISA, for instance, utilises two basic procedures: indirect ELISA, where an antiserum is used to trap a particular IgG [197], and direct ELISA, which relies solely on the antibody-enzyme conjugate to attach to host-specific IgG in the bloodmeal [68,198], currently preferred for its simplicity over indirect ELISA. PCR, being more sensitive due to specific primers targeting host DNA, has evolved from conventional PCR, which amplified human host DNA extracts at the human tyrosine hydrolase (TC-11 or HUMTHO1) and VWA (HUMVWFA31) [199,200], to the current multiplexed PCR capable of detecting five mammalian blood meals in mosquitoes in a single step (i.e., by size-differentiated DNA fragment on agarose gels) [69]. While these techniques offer significant advantages, they also come with challenges such as being time-consuming, laborious, and require repeated use of expensive reagents, not always readily available in rural laboratories where field collections are conducted. Moreover, ELISA assays, one of the most widely used technique, are prone to high levels of cross-reactivity, occasionally failing to sufficiently distinguish between human and hon-human blood meals [74]. Since field collections do not always yield synchronous physiological states, some of the blood meals may have been partly digested, which might also compound the detection capability of current methods [201].

In a recent study, our team demonstrated that machine learning models trained on mid-infrared spectra data collected from mosquitoes fed on different hosts (4000 cm^{-1} to 400 cm^{-1} frequencies) (MIRS-ML) could accurately distinguish vertebrate blood meals in laboratory-reared *An. arabiensis* mosquitoes without the need for molecular techniques [114]. However, it was also noted that field validation would be necessary for multiple reasons. Firstly, in field settings, the time post-feeding is unknown, and the mosquitoes may have multiple blood meals, occasionally from multiple sources. Secondly, unlike laboratory settings where the age of mosquitoes is known, field mosquitoes vary in age and may have taken their 2nd, 3rd, or 4th meals. Thirdly, the amount of blood in the mosquito gut may be small in the field due to increased disturbance during feeding compared to controlled laboratory conditions, and lastly, the genetic variability for blood sources is higher in the field. Overcoming these challenges would enable the potential use of MIRS-ML in real-world field scenarios. We, therefore, concluded from the initial laboratory study that whereas the technique offers a unique opportunity to rapidly test individual mosquitoes for blood-type and other attributes, assessing blood-feeding histories of wild malaria mosquitoes would provide an opportunity to test its potential field validation.

The current study aimed to analyse the blood-feeding preferences of wild-caught malaria mosquitoes, by using MIRS-ML models to identify the sources of their blood

meals. The study also explored how well the models trained using laboratory-reared mosquitoes can be applied to field-collected samples by incorporating specific transfer learning techniques previously used for predicting the species identification and age of mosquitoes collected in different countries [61,178]. The ultimate goal of the work was to demonstrate the utility of this approach for field applications. Implementing these models in the field would significantly enhance the knowledge of mosquito feeding behaviours and disease transmission, potentially informing more effective vector control strategies against multiple mosquito-borne diseases [18,43–49].

4.3 Methods

4.3.1 Mosquito collection and processing

Mosquitoes were sampled from five sites in Tulizamoyo, a rural village in Ulanga district, southeastern Tanzania (8.3544°S, 36.7054°E). To capture a comprehensive range of blood-meals, collections were conducted as follows: a) indoors using CDC light traps and resting buckets throughout the night (6:30 PM to 6:30 AM) and Prokopack aspirators during the early morning (5:30 AM to 6:30 AM); b) outdoors in peri-domestic areas, including outdoor kitchens, with the same night and early morning methods; and c) around animal sheds, again using resting buckets at night and Prokopack aspirators in the morning.

The collected mosquitoes were sorted by taxa and physiological states [202]. All blood-fed *Anopheles* females were killed with chloroform and preserved individually in 1.5 mL Eppendorf tubes containing silica gel desiccant afterwards. The mosquitoes were kept for 5 days at 5°C before scanning (see below). In total, 1,854 blood-fed (76% *An. funestus* and 24% *An. arabiensis*) females were examined.

4.3.2 Mid-infrared spectrometer scanning

The abdomens of all blood-fed *An. funestus* and *An. arabiensis* were scanned. An attenuated total reflection Fourier-transform infrared (ATR-FTIR) ALPHA II spectrometer (Bruker optics) was used to collect the infrared spectra of dried mosquito abdomens over a spectral range of 4,000 to 400 cm^{-1}, with a 2 cm^{-1} resolution. The absorbance data obtained from scanning the mosquito abdomens provides insights into the biochemical makeup, e.g. the protein and lipid concentrations present in the blood meal, which are indicative of the vertebrate source of the blood meal [114]. Each mosquito was scanned 32 times and the

spectra were averaged. Scanning was done inside the Ifakara Health institute's Vector Biology Laboratory, the VectorSphere.

4.3.3 Identification of blood meals from different vertebrate hosts using polymerase chain reaction (PCR)

Following MIRS analysis, mosquito carcasses were subjected to a multiplex PCR assay to identify the vertebrate origins of their blood meals as either from humans, cows, goats, dogs, or pigs. A multiplexed PCR assay was used targeting the cytochrome b (cytB) gene following the Kent *et al.*, protocol [69]. DNA was extracted using DNAzol® with a final volume of 20μL per sample. The PCR mix included 5μL of DNA, 1μL each of 20μM universal and species-specific primers, and 12.5μL of One Taq Quick Load 2X master mix. Amplification conditions were: 95°C for 5 minutes, 29 cycles of 95°C for 1 minute, 58°C for 1 minute, 72°C for 1 minute, and a final extension at 72°C for 7 minutes. The PCR products were run on a 2% agarose gel with Classic view stain and imaged under UV light with the Kodak Logic 100 system, assessed in comparison to the known fragment sizes for different hosts (Kent *et al.*, [69] as shown in Table 4.1). PCR results were used as the "ground truth" to train and validate machine learning algorithms.

Table 4.1: Amplified DNA fragments from different blood meal hosts

Host Blood	Fragment size (base pairs)
Human	334
Bovine	561
Goat	132
Dog	680
Pig	453

4.3.4 Confirmation of the identity of sibling species in the *An. funestus* group

Using DNA extracted from the same mosquitoes, a multiplex PCR protocol by Koekemoer *et al.*, [180] was used to identify and distinguish between sibling species within the *An. funestus* group.

4.3.5 Training machine learning models to identify and distinguish between blood meal types

The analysis was carried out in Python 3.9 using the Scikit-learn [141] and Keras [145] libraries for the machine learning tasks. Supervised machine learning was exclusively trained with wild-caught *An. funestus* females dataset (N = 751), consisting of human-blood fed ($n = 167$) and bovine blood-fed ($n = 584$) mosquitoes, in order to predict blood meal sources for field-collected mosquitoes. Before performing model training and prediction, the classes were balanced by randomly under-sampling the over-represented blood meal class to match the under-represented classes (i.e., human-blood fed ($n = 167$) and bovine blood-fed ($n = 167$) mosquitoes). The remaining samples from the random under-sampling were later included in the unseen data/test data for overall prediction. Field collected *An. arabiensis* were not used for model training since there were only 256 (human blood-fed ($n = 2$) and bovine blood-fed ($n = 254$)) of them in the total sample set. Additionally, prior to model training, the spectra were cleaned of water vapour absorption bands and carbon dioxide (CO_2) interference bands then standardised by rescaling to zero mean and a variance of 1 to ensure consistency and uniformity. The following algorithms were tested and compared to select the one with the highest predictive accuracy and precision: K-nearest neighbours (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (XGB), Random Forest (RF), and Multilayer Perceptron (MLP). The best-performing model was selected based on predictive accuracy and refined it through hyper-parameter tuning. This optimised model was then validated using 5-fold cross-validation. Once the model was validated, it was tested using a balanced set of unseen spectra from human blood-fed ($n = 17$) and bovine blood-fed ($n = 17$) mosquito samples derived from the under-sampling process.

A second-stage model evaluation was conducted using a larger but imbalanced set of test samples consisting predominantly of spectra from bovine-fed mosquitoes ($n = 688$) and a small number of spectra from human-fed mosquitoes ($n = 19$). While the datasets used for both the model training and the first stage testing consisted of only *An. funestus*, this larger dataset used for the second stage testing also included a small number of blood-fed *An. arabiensis* ($n = 254$), which had been excluded from model training.

Lastly, a transfer learning technique was implemented to predict field data by initially training machine learning models with laboratory data and then augmenting with small quantities of field data as follows. In this context, deep learning framework was utilised due to their direct provision of pre-trained models and pre-build transfer learning capabilities, which differs from traditional machine learning algorithms. Spectral data from a previous study were utilised [114], which involved laboratory-reared mosquitoes to train the deep learning model. This earlier study used age-synchronised lab-reared *An. arabiensis* fed

on four different host types, cattle, goat, chicken and humans [114]. This pre-existing data was used here to train an MLP deep learning model within the Keras framework, but only the mosquitoes fed on human blood ($n = 409$) and bovine blood ($n = 454$) were included. Then, the model was augmented with a small subset of newly collected data from wild mosquitoes to assess the amount of field data needed for effective transfer learning. The resulting MLP model was then utilised to classify the sources of blood meals in wild-collected mosquitoes from two different test sets: a near-balanced set of test samples (human blood-fed ($n = 177$) and bovine blood-fed ($n = 120$)) derived from the under-sampling process, and an imbalanced set of test samples consisting predominantly of spectra from bovine-fed mosquitoes and a small number of spectra from human-fed mosquitoes; the second test set included 784 bovine blood-fed and 122 human blood-fed mosquito samples.

While accuracy was the primary evaluation metric for the model, additional metrics, namely recall (true positive rate), precision (positive predictive value), and F1-scores were also employed for a comprehensive performance assessment. The recall score, indicating the ability of the model to identify all actual positives and minimise false negatives, was calculated as the proportion of accurately identified blood meal hosts out of the total blood-fed mosquitoes within each category. Precision, reflecting the success of the model in avoiding false positives, was measured as the proportion of correctly classified blood meal host/source against all the positive predictions of that model for each blood meal category. Lastly, the F1-score, a harmonic mean of precision and recall, was computed to gauge the balanced performance of the model in accurately classifying blood meal host sources. A higher F1-score denotes superior model efficacy, with a score of 1 indicating perfect precision and recall.

4.3.6 Estimating the human blood index (HBI) from polymerase chain reaction (PCR) and mid-infrared spectroscopy and machine learning (MIRS-ML) approaches

The proportion of mosquito blood meals obtained from humans were estimated through predictions generated by MIRS-ML based approaches and compared them to the outcomes of PCR analysis. The definitive 'ground truth' HBI (human-fed/total blood-fed mosquitoes) was calculated using PCR results, while MIRS-ML based prediction were used for comparison.

4.4 Results

4.4.1 Polymerase chain reaction (PCR) based identification of blood meals from different vertebrate hosts

A total of 1,854 samples were examined (Table 4.2). Of these 45.2% of the mosquitoes had consumed bovine blood, 9% human blood, 3.7% dog blood, and 1.4% a mixture of human and bovine blood. Another 0.3% had fed on either a mix of human and dog blood or bovine and dog blood. Notably, 40.1% of all samples remained unamplified, possibly due to prolonged host-blood digestion within the mosquito abdomen [201] or the presence of blood from other vertebrates not targeted by the list of primers used in the study.

Table 4.2: Number of amplified host blood meal sources of wild-caught *Anopheles* mosquitoes

Host Blood	***An. funestus* group**	***An. arabiensis***	**Total Count (%)**
Bovine blood	584	254	838 (45.2)
Human blood	167	2	169 (9)
Dog blood	65	3	68 (3.7)
Human & bovine blood	26	-	26 (1.4)
Bovine & dog blood	5	-	5 (0.3)
Human & dog blood	5	-	5 (0.3)
Unamplified	553	190	743 (40.1)
Total	**1,405**	**449**	**1,854 (100)**

4.4.2 Confirmation of the identity of sibling species in the *Anopheles funestus* group

Additional PCR was conducted to determine the species composition of *An. funestus* that blood-fed on bovine and humans. These tests revealed that 99% of the successfully amplified bovine blood-fed samples were *An. funestus*, with *Anopheles rivolurum* and *Anopheles vaneedeni* making up 0.7% and 0.1%, respectively. *An. funestus* also accounted for 100% of the amplified samples from mosquitoes that had fed on human blood.

4.4.3 Using machine learning models to identify and distinguish between blood meal types

As humans and cattle were found to be the predominant hosts (Table 4.2), the ML models were exclusively trained using labels from *An. funestus* human blood-fed (n = 167) and bovine blood-fed (n = 584). To address the imbalance, the bovine blood-fed class was under-sampled at random to match the under-represented class (i.e., human-blood fed (n = 167) and bovine blood-fed (n = 167) mosquitoes) [203].

LR achieved the highest in-sample prediction accuracy at 80% (Fig. 4.1A). After hyper-parameter tuning, the LR model predicted the previously unseen balanced set of test samples with an overall accuracy of 88%, with 94% for bovine and 82% for human blood meal classifications (Fig. 4.1B). The summarization of this result on a confusion matrix shows that about 6% of mosquitoes blood-fed on bovine were misclassified as human blood-fed, and 18% of human blood-fed mosquitoes were misclassified as bovine blood-fed (Fig. 4.1B).

Moreover, when all the remaining samples were included in the test set to create a larger but imbalanced dataset, the LR model classified all the previously unseen spectra with an overall accuracy of 78%, predicting bovine blood-fed and human blood-fed mosquitoes with 73% and 82% accuracy, respectively (Fig. 4.1C). Additionally, a lower precision was observed for the minority class (i.e. Human). Additional metrics (precision, recall and F1 statistics) and the number of test samples are in Table 4.3.

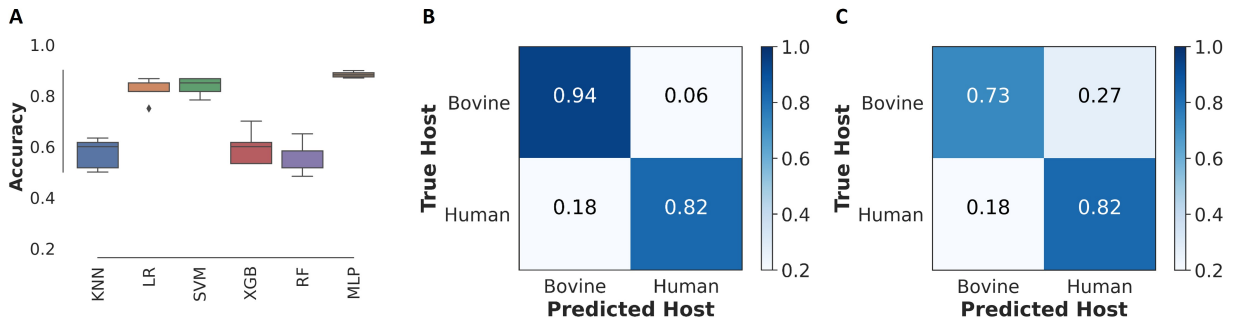

Figure 4.1: A) Comparison of machine learning algorithms; K-nearest neighbours (KNN), Logistic regression (LR), Support vector machine (SVM), Extreme Gradient boosting (XGB), and Random forest (RF). **B**) A confusion matrix from the LR classifier's predictions on the balanced set of test samples of wild *An. funestus* blood-fed on human and bovine. **C**) A confusion matrix from the LR classifier's predictions of the imbalanced set of test samples of wild mosquitoes blood-fed on human and bovine.

Table 4.3: Precision, recall, and F1-score of the LR classifier in classifying Bovine and human blood-meal sources in out-of-sample wild malaria mosquitoes

Host blood	Precision	Recall	F1-score	No. test samples
Model testing using a balanced set of test samples				
Bovine	0.84	0.94	0.89	17
Human	0.93	0.82	0.87	17
Model testing using an imbalanced set of test samples				
Bovine	0.99	0.79	0.88	688
Human	0.09	0.79	0.17	19

4.4.4 Using machine learning models trained with laboratory data to classify host blood meals of field-collected mosquitoes

Although the initial model trained with field data yielded a relative high accuracy performance, the effectiveness of a model trained using laboratory data from an earlier study was evaluated [114], for classifying the host blood meals of field-collected samples. Indeed, the advantage of this approach is that it would allow to create models using laboratory samples, which are easier to produce and balance between different hosts.

After training a baseline MLP model, a small subset of field spectra was incorporated using transfer learning which can allow generalisation with minimal re-calibrations [61]. Transfer learning exhibited a significant enhancement in classification accuracy, increasing from 76% to approximately 90% (Fig. 4.2A). This level of accuracy was achieved by integrating, into the MLP model trained with laboratory data up to 100 field samples, evenly split between human-fed and bovine-fed classes. Specifically, on the balanced set of test samples, the MLP model achieved a classification accuracy of 90% for bovine blood meal sources and 91% for human blood meal sources (Fig. 4.2B).

Moreover, on the imbalanced set of test samples (784 bovine blood-fed and 122 human blood-fed), the MLP model improved and achieved an overall accuracy of 94%, with 98% for bovine and 90% for human blood-fed mosquitoes (Fig. 4.2C). The precision, recall, F1-score metrics, and the number of test samples are presented in Table 4.4.

Lastly, to assess whether MIRS-ML could be used to estimate human blood index (HBI), which reflects the proportion of mosquito blood meals derived from humans, the predictions by MIRS-ML were compared against standard HBI values obtained by PCR. It was observed that LR predictions, when solely based on field data, slightly underestimated the HBI by 6% compared to PCR results. On the other hand, the predictions obtained by the model that used transfer learning were much more accurate in estimating HBI; and

even minimal number of samples included in the re-calibration model well aligned with the PCR-based standard HBI (Fig. 4.3).

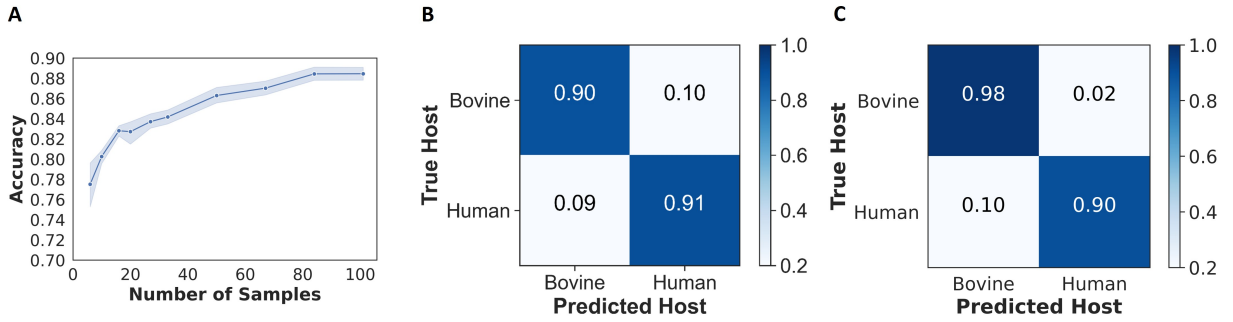

Figure 4.2: **A**) The accuracy of classifying unseen blood-meal sources in field mosquitoes significantly increased from 76% to 90% when using a training set of up to 100 field mosquitoes for transfer learning. The mean accuracy is depicted by the solid line, while the shaded ribbon represents the standard deviation of the mean across 10 models. **B**) A confusion matrix from the transfer learning model for classifying human and bovine blood meals in field mosquitoes from the balanced set of test samples. **C**) A confusion matrix from the transfer learning model's classification prediction of the imbalanced set of test samples of wild mosquitoes blood-fed on human and bovine.

Table 4.4: Precision, recall, and F1-score of the transfer learning model (i.e. MLP) in classifying out-of-sample bovine and human blood-meal sources in wild malaria mosquitoes.

Host blood	Precision	Recall	F1-score	No. test samples
Model testing using a balanced set of test samples				
Bovine	0.91	0.90	0.90	120
Human	0.90	0.91	0.90	117
Model testing using an imbalanced set of test samples				
Bovine	0.98	0.98	0.98	784
Human	0.88	0.90	0.89	122

4.5 Discussion

Human blood index (HBI), which reflects the tendency of mosquitoes to feed on humans compared to other vertebrates, is vital for assessing malaria transmission dynamics and the level of stability of transmission [41]. Current techniques for determining mosquito blood meal sources are slow, labour-intensive, and expensive due to the need for costly reagents. They are also susceptible to errors, such as false positives from cross-reactivity with other antigens or due to the partial digestion of blood meals in the mosquito digestive system. Yet, as malaria endemic countries move towards elimination, there is a pressing need for simpler, more cost-effective methods that can be deployed at scale in malaria-endemic countries to improve entomological surveillance and evaluate the effectiveness of malaria control interventions.

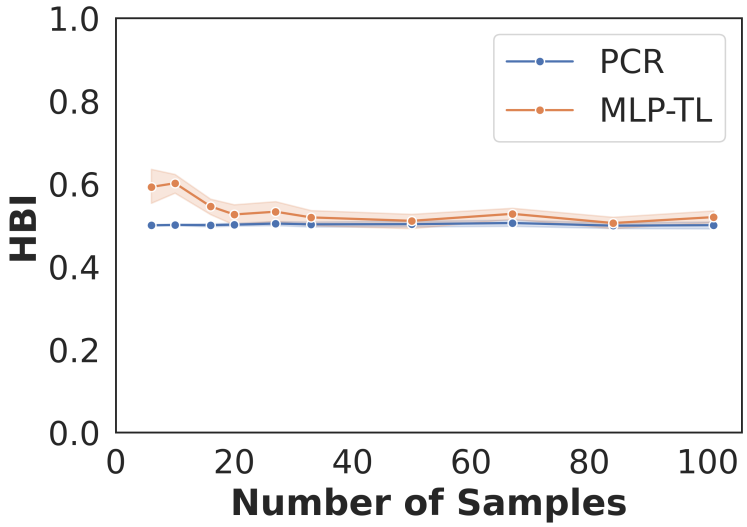

Figure 4.3: Estimation of the HBI by the transfer learning (i.e. MLP-TL, Multilayer perceptron-transfer learning) compared to PCR when using a training set of up to 100 field mosquitoes for transfer learning. The solid line represents the average HBI, while the shaded ribbon illustrates the standard deviation across 10 iterations.

This study demonstrates the first-ever field application of the simple mid-infrared spectroscopy and machine learning (MIRS-ML) approach for predicting the blood-feeding histories of malaria vector in rural Africa. Beyond this, the study also demonstrates the transferability of the laboratory-trained MIRS-ML models to identify and classify host blood meals in field-collected samples through the utilisation of transfer learning techniques. For validation, PCR as the 'ground truth' was used to determine the actual blood-feeding histories of the field-collected mosquitoes; and examined a total of 1,854 blood-fed *Anopheles* mosquitoes.

Based on the PCR analysis, most of the mosquitoes blood-fed on humans or bovines, and only a very small percentage had fed on other hosts, such as dogs and pigs. Given the inherent limitations of the PCR, classification of blood meals in 41% of the samples was impossible, possibly because they fed on a host other than those tested in this study and therefore could not be amplified with the primers used. Nonetheless, only mosquitoes confirmed to have fed on either humans or bovines were included in this analysis, as they were the vast majority; thus binary machine learning classifiers were trained for blood-meal prediction. The capability of the MIRS-ML models to classify mosquito blood-meal sources was demonstrated, achieving an accuracy of 88%, when using 338 spectra data collected from field samples (169 human-fed and 169 from bovine-fed mosquitoes). This demonstrates a realistic opportunity to deploy such simple methods for estimating HBI, thereby extending the capability of infrared-AI based systems already well demonstrated for tracking several other entomological attributes [204].

In prior work using age-synchronised laboratory-reared mosquitoes, the focus was on predicting blood-meal sources with *An. arabiensis*, where the MIRS-ML approach

achieved a classification accuracy of ~98% for four blood meal sources (bovine, human, goat and chicken) [114]. Whereas the mosquitoes used in that earlier study were only 6 to 8 hours post-feeding, this current study included a broader range of age groups and natural variation in the degrees of digestion of the blood meals. This current study therefore strongly demonstrates the potential of the MIRS-ML approach for realistic field surveillance, even when the time of actual blood-feeding and digestion stages is unknown upon sample collection and preparation.

A major achievement in the present work is the demonstration of the transferability of laboratory-trained models to field samples through the application of transfer learning. The transferability of laboratory-trained models achieved a classification accuracy of 90% in predicting blood-meal sources for field-collected *An. funestus*. The base laboratory model was initially trained using spectra data from blood-fed *An. arabiensis* [114], which was then augmented by incorporating a small subset ($n = 100$, with 50 samples each from humans and bovine blood-fed *An. funestus* spectra) of field-collected data into the model. This implies that the technique can be extended to assess blood-meal sources in the abdomens of Afro-tropical malaria vectors, as the species would not be a confounding factor in this case. It also implies that the generalisability of this model will cut across laboratory and field sample prediction, and therefore, sample origin might not be a confounding factor. Since field-collected mosquitoes were likely of varied ages, and therefore mosquito age, a factor readily classifiable by MIRS-ML models [61], is also unlikely to be a confounder, and can be overcome by similar transfer learning approaches. The results presented here corroborate with previous studies in which the utilisation of transfer learning successfully generalised predictions of mosquito age and species across different countries and laboratories [61,178]. This approach effectively accounts for the inherent variability of mosquitoes from different environmental and ecological settings or genetic backgrounds, which could otherwise limit the generalisability of ML models trained on mosquito spectra data to new mosquito populations. Indeed, the genetic variability for blood meals in the field is likely high, and blood-fed mosquitoes collected during the study contained a mixture of fully engorged and partially consumed blood meals.

Partial digestion or low quantity of ingested blood meals, could potentially impair the capability of MIRS-ML to accurately identify or differentiate between various blood meals, thereby affecting the Human Blood Index (HBI) estimates. To mitigate this, it is advised against including gravid mosquitoes in samples and recommended to preserve all blood-fed mosquitoes immediately upon collection to halt any biochemical changes before spectroscopy. Currently investigating this phenomenon, preliminary studies have demonstrate a notable decrease in MIRS-ML accuracy after 36 hours post-feeding (Mgaya *et al.*, (unpublished), which coincides with gravidity in a typical 2-3 day gestation period under optimal conditions. In this paper, field models closely aligned with PCR outcomes, considered as the benchmark, despite the inability to precisely determine the gestational

stage of mosquitoes at the time of collection each morning post-trapping. Moreover, earlier studies by Mukabana *et al.*, [201], have successfully used PCR to amplify host DNA up to 32 hours post-feeding after which the host DNA is degraded. Crucially, the analysis only incorporated samples that yielded successful PCR amplification of host DNA for MIRS-ML training, discarding all non-amplified samples. This selection criterion may inadvertently introduce bias since the partially or fully digested blood meals may be the ones least likely to yield good-quality host DNA. Future models should therefore include samples of mosquitoes that have blood-fed on known hosts, 1-4 days post-feeding to evaluate the efficacy of MIRS-ML across various stages, including gravid and post-oviposition states. Lastly, though the model was already trained on a large number of mosquitoes, it is recommended to increase these sample sizes and obtain mosquitoes from different sampling locations so as to neutralise effects such as partial blood-meals and partial digestion, as well as any effects of environmental or micro-climatic factors affecting blood feeding and digestion.

Indeed, increasing the number of field samples for transfer learning not only enhanced the classification accuracy for field blood-fed mosquitoes but also improved the precision in estimating the HBI in comparison to the 'ground truth' PCR method. This indicates that the technique has the potential to be a reliable method for estimating HBI, capable of generalising HBI estimations in field-collected mosquitoes as effective as PCR. Therefore, it can provide valuable information to national malaria control programs regarding the feeding preferences of malaria mosquitoes.

Despite the successes of this technique, there remain several gaps. Firstly, it is unclear whether the technique can detect mixed blood meals, a situation that is more likely to occur in the field, remains unanswered, warranting future investigation. Secondly, PCR and ELISA remains highly sensitive and specific, known for their accuracy in detecting host DNA and specific protein from blood meals, even in small amounts, respectively. Although MIRS-ML has demonstrated notable accuracies in detecting mosquito blood meals, its performance, being highly sensitive and specific, depends on the quantity and quality of the training data and machine learning algorithms used. This robustness of the model will contribute to its ability to handle variations. Thirdly, the machine learning models in this study were trained using *An. funestus* mosquitoes that had blood-fed on humans and bovines. This choice was made because most mosquito samples collected from the field contained either human or bovine blood in their abdomens, while only a minority had dog blood or mixed blood-meals. Consequently, the available samples were insufficient to adequately train the machine learning models to detect mosquito blood-meal sources from hosts other than humans and bovines. In their current state, these models would face challenges in field deployment since they will not be capable of identifying blood-meal sources from other potential hosts often found in human dwellings such as goat, pig, and chicken. However, considering that the transferability of the laboratory-trained

models for field sample prediction has also been demonstrated, the deployment of these models could involve initially training them on laboratory data, which can be generated in large quantities. Additionally, this approach allows for the inclusion of a wider range of hosts, ensuring accurate mosquito blood-meal source prediction from all common hosts typically found near human dwellings, including humans, bovines, goats, dogs, pigs and chickens. Thus, once validated, MIRS-ML approaches have the potential to make significant contributions to understanding the dynamics of disease transmission involving humans, livestock, wildlife, and vectors. Specifically, they could offer valuable insights into scenarios where mosquitoes have opportunities to feed on multiple host species.

Interestingly, despite its anthropophilic behaviour, *An. funestus*, the main vector in the study area, was found to also blood-feed on bovines. This finding is consistent with previous studies that demonstrated a potential switch in host choice by *An. funestus* from humans to cattle [205, 206]. In brief, given the circumstances of the collections, this observation may be explained by several factors: Firstly, the houses where mosquito collections were conducted had been supplied with intact bed nets before the collections started, which might have created a physical barrier, reducing mosquito exposure to humans [64]; and forcing mosquitoes to use alternative blood sources in the surrounding areas as previously documented by Iwashita *et al*, [64]. Secondly, it might have been a result of the zoopotentiation effect, which refers to the increased tendency of mosquitoes to feed on humans living near livestock [65,66], especially when livestock in close proximity to human dwellings emit heat and odour cues that attract mosquitoes. In such circumstances, not only do zoophagic mosquitoes find additional blood sources that they already prefer, but even the naturally anthropophagic mosquitoes may also accidentally feed on cattle when host cues become mixed nearby. There is a lot of evidence suggesting that zoopotentiation may increase malaria transmission risk by creating an alternative source of bloodmeals, consequently increasing both mosquito survival rates and abundance [67, 207–210]. This interaction of mosquitoes between humans and non-human hosts may also elevate the likelihood of transmitting parasitic helminths and zoonotic pathogens [18, 43–49, 211].

Infrared spectroscopy and machine learning methods have already been demonstrated for several other use cases, such as age-grading mosquitoes [59–61, 94, 178], detection of pathogens inside mosquitoes [93], identification of mosquito species [61] and even detection of parasites in human blood [113, 115, 126]. This demonstration of its usefulness for analysing the blood-feeding histories of mosquitoes in both the laboratory (as previously shown [114]) and the field (this current study), underscores the unique potential of the technology as a one-stop system for comprehensive analysis of entomological and parasitological indicators of malaria and other mosquito-borne diseases.

4.6 Conclusion

In conclusion, the study marks the pioneering application of mid-infrared spectroscopy combined with machine learning (MIRS-ML) for the rapid assessment of blood-feeding patterns in field-collected malaria vectors. By successfully classifying the blood meals of wild *An. funestus* female mosquitoes, it has been demonstrated that, regardless of whether the ML models were trained with MIR spectra from field-collected conspecific females or from laboratory-reared *An. arabiensis*, MIRS-ML has the accuracy, precision, and overall potential for identifying and distinguishing between different host blood meals. By comparing results with multiplex PCR assays, considered the 'ground truth', MIRS-ML achieved high classification accuracies of 88% and 90% with logistic regression and multi-layer perceptron classifiers, respectively. Notably, the study also confirms the effectiveness of transfer learning in adapting laboratory-trained models for field data analysis. The MIRS-ML method represents a scalable, cost-efficient alternative to traditional, more labour-intensive blood meal analysis methods, and has the added advantage of estimating the human blood index (HBI) with only slight overestimation. Since this technology has already been demonstrated for several other entomological and parasitological surveys, this study demonstrates its extended capability and potential as a 'one-stop' system for comprehensive analysis of entomological and parasitological indicators of malaria and other mosquito-borne diseases. This advancement is crucial for malaria-endemic regions seeking simpler analytical methods to enhance entomological surveillance or to evaluate the impact of disease control efforts. The marginal discrepancies in HBI estimation do not detract from the method's utility; rather, they highlight the transformative potential of MIRS-ML in facilitating comprehensive surveillance and providing deeper insights into malaria transmission dynamics.

4.7 Ethics approval and consent to participate

Ethical approval for this study was obtained from Ifakara Health Institute Institutional Review Board (Ref. IHI/IRB/No: 41-2020), and from the Medical Research Coordinating Committee (MRCC) at the National Institute of Medical Research (NIMR), Ref: NIMR/HQ/R.8a/Vol. IX/3557.

4.8 Code, data, and materials availability

The mid-infrared spectral datasets generated and analysed during the current study, as well as code for the analyses is available at https://github.com/MwangaEP/Field-bloodmeal-MIRS.

4.9 Competing interests

The authors declare no competing interest.

4.10 Funding

This study was supported by the Wellcome Trust Masters Fellowship in Tropical Medicine & Hygiene (Grant No. 214643/Z/18/Z) awarded to EPM and the Medical Research Council (MRC) [MR/P025501/1] awarded to FB. FB is supported by the Academy Medical Sciences Springboard Award (ref:SBF007/100094). FOO was supported by a Howard Hughes Medical Institute (HHMI)-Gates International Research Scholarship (Grant No. OPP1099295), and Bill and Melinda Gates foundation (INV003079). SAB is supported by the Bill and Melinda Gates Foundation (INV-030025) and Royal Society (ICA/R1/191238).

4.11 Authors' contributions

EPM, SB, FB, KW, PS, MTS and FOO conceived the study. EPM, SA, FOO, and FB developed the study's protocol. GS, FEM and EPM collected the data. ISM, SA, and EPM performed molecular assays. EPM carried out data analysis and ML training. EPM wrote the manuscript. EPM, DJS, SHM,NJM, IHM, MGJ, KW, MTS, PS, SB, FB and FOO reviewed and edited drafts of the manuscript. All authors have read and approved the final manuscript.

Chapter 5

Reagent-free Detection of *Plasmodium falciparum* Malaria Infections in Field-collected Mosquitoes Using Mid-infrared Spectroscopy and Machine Learning

Emmanuel P. Mwanga[1,2*], Prisca A. Kweyamba[1,3,4], Doreen J. Siria[1,2], Issa H. Mshani[1,2], Idrisa S. Mchola[1], Faraja E. Makala[1], Godian Seleman[1], Said Abbasi[1], Sophia H. Mwinyi[1], Mario González-Jiménez[5], Klaas Wynne[5], Francesco Baldini[1,2], Simon A. Babayan[2], Fredros O. Okumu[1,2,6,7].

1. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, Morogoro, Tanzania.
2. School of Biodiversity, One Health and veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, UK.
3. Swiss Tropical and Public Health Institute, Kreuzstrasse 2, CH-4123 Allschwil, Switzerland.
4. University of Basel, Petersplatz 1, CH-4001 Basel, Switzerland.
5. School of Chemistry, The University of Glasgow, Glasgow, G12 8QQ, UK.
6. School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

DOI: https://doi.org/10.1038/s41598-024-63082-z

5.1 Abstract

Field-derived metrics are critical for effective control of malaria, particularly in sub-Saharan Africa where the disease kills over half a million people yearly. One key metric is entomological inoculation rate, a direct measure of transmission intensities, computed as a product of human biting rates and prevalence of *Plasmodium* sporozoites in

mosquitoes. Unfortunately, current methods for identifying infectious mosquitoes are laborious, time-consuming, and may require expensive reagents that are not always readily available. Here, we demonstrate the first field-application of mid-infrared spectroscopy and machine learning (MIRS-ML) to swiftly and accurately detect *Plasmodium falciparum* sporozoites in wild-caught *Anopheles funestus*, a major Afro-tropical malaria vector, without requiring any laboratory reagents. We collected 7,178 female *An. funestus* from rural Tanzanian households using CDC-light traps, then desiccated and scanned their heads and thoraces using an FT-IR spectrometer. The sporozoite infections were confirmed using enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR), to establish references for training supervised algorithms. The XGBoost model was used to detect sporozoite-infectious specimen, accurately predicting ELISA and PCR outcomes with 92% and 93% accuracies respectively. These findings suggest that MIRS-ML can rapidly detect *P. falciparum* in field-collected *An. funestus*, with potential for enhancing surveillance in malaria-endemic regions. The technique is both fast, scanning 60-100 mosquitoes per hour, and cost-efficient, requiring no biochemical reactions and therefore no reagents. Given its previously proven capability in monitoring key entomological indicators like mosquito age, human blood index, and identities of vector species, we conclude that MIRS-ML could constitute a low-cost multi-functional toolkit for monitoring malaria risk and evaluating interventions.

5.2 Background

Vector surveillance is an essential component of malaria control and elimination, and generally includes an assessment of prevailing transmission intensities, the behaviours of different vector species and the responsiveness of these species to different interventions [32]. The most direct metric of malaria transmission intensities is the entomological inoculation rate (EIR), which is the number of infectious bites per person in a unit time, and is defined as the product of the human biting rate (HBR) and proportion of the biting mosquitoes that have *Plasmodium* sporozoite in their salivary glands [33–35]. While other entomological parameters such as mosquito abundance, age structure, daily survival probabilities, larval densities and blood-feeding preferences are important, EIR is also used to estimate the level of exposure and analyse the effectiveness of control programs. However, current reports indicate that not all endemic countries possess transmission intensity data or measure sporozoite rates [34, 212]. Arguably, therefore, having a simpler method for testing samples might improve these surveillance capabilities in endemic countries.

Plasmodium infections in mosquitoes can be detected using various techniques, the main ones being enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction

(PCR), which are both used widely, especially in research settings [75, 78, 79, 213, 214]. Other more traditional approaches include dissection and microscopic examination of the salivary glands 8 and Loop-Mediated Isothermal Amplification (LAMP) assays [76]. These techniques, despite being key features in many laboratories, present several challenges, which often limit their adoption for programmatic use beyond research projects. For example false positivity rates have been reported in ELISA assays, especially where malaria vector species with zoonotic behaviours are screened, in which cases a number of non-target protozoans may be picked up in the assays, potentially leading to an overestimation of EIR [84, 85]. More importantly, despite their benefits attained by both PCR and ELISA, PCR is generally expensive due to the cost of reagents. Moreover, the reagents for both PCR and ELISA are often not readily available in the localities where they are most needed. They are also time-consuming, requiring significant efforts and specialised laboratory facilities for sample preparation and processing [60, 93]. Lastly, all the methods, including hand dissections of the salivary glands require highly trained and experienced personnel. These challenges underscore the critical necessity for innovative approaches that not only achieve high accuracy in detecting malaria parasites in mosquitoes but are also cost-effective, rapid, and user-friendly. Such a system would be beneficial in low-income, malaria-endemic countries, where the WHO's recommendation to incorporate surveillance as a fundamental pillar of malaria programs [32] is hindered by the absence of easily scalable systems for effective surveillance.

Recently, the use of infrared spectroscopy, specifically near-infrared spectroscopy (NIRS, 12,500 - 4,000 cm^{-1} frequencies of the electromagnetic spectrum), has shown potential for detecting the presence of *Plasmodium spp.* in *Anopheles* mosquitoes under controlled laboratory settings [93, 215]. However, in a field validation of this technique, the predictive models could not distinguish between sporozoite-infectious and non-infectious mosquitoes [92]. Mid-infrared spectroscopy (MIRS), which uses frequencies between 4000 - 400 cm^{-1}, can provide clearer peaks with more detailed information than NIRS [98, 174], and has been hypothesized to carry greater potential for such applications. Advancements in machine learning and deep learning algorithms are enhancing the potential of spectroscopic data analysis by enabling more detailed examination. This advancement allows for better specimen classification and a more detailed understanding of how samples differ in their biochemical composition [59, 61, 114, 115, 178, 216].

By integrating MIRS spectroscopic techniques and machine learning approaches, it has been possible to measure multiple entomological and parasitological indicators of malaria transmission. Examples include identifying epidemiologically relevant species and age groups of *Anopheles* mosquitoes [59, 61, 178], evaluating the blood-feeding histories of mosquitoes to determine preferences for either humans or other vertebrates [114], and detecting *Plasmodium falciparum* infections in human blood samples collected from malaria endemic villages [113, 115, 126, 216]. However, the ability of MIRS to detect natural

Plasmodium infections in wild-caught malaria vectors has not been demonstrated, a capability which is greatly needed to estimate malaria transmission intensities in endemic settings.

This current study was therefore designed to demonstrate the first field application of mid-infrared spectroscopy combined with machine learning (MIRS-ML) for rapid and accurate detection of *P. falciparum* in field-collected *An. funestus*. To achieve this, we evaluated the technique using PCR and ELISA as the 'ground truth' to detect *P. falciparum* sporozoites in wild-caught *An. funestus*, the leading malaria vector in Tanzania [28,128,130].

5.3 Results

5.3.1 Prevalence of *P. falciparum* sporozoites in *An. funestus* as detected by enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR)

The ELISA screening detected 184 positives out of the 4281 tested samples (4%) while the PCR screening method detected 144 positives out of the 2897 tested samples (5%).

5.3.2 Machine learning classifications of mid infrared spectra of infectious and non-infectious *An. funestus*

To differentiate between infectious and non-infectious *An. funestus* mosquito spectra (refer to Fig. 5.1A), four of the six machine learning models we tested achieved prediction accuracies above 85% (Fig. 5.1B). Prediction accuracy refers to the proportion of correct predictions (both true positives and true negatives) made by a model out of total predictions. XGBoost was selected for further tuning of the model settings with the aim of finding the optimal combination of parameters for improved performance. This choice was made due to the capability of the XGBoost model to capture relationships between variables in the data, particularly those that do not follow straight line or a simple curve [217].

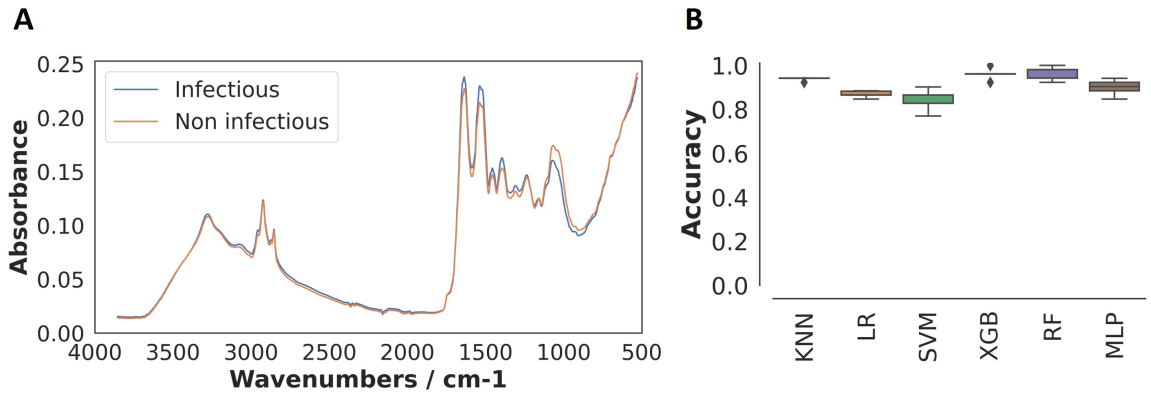

Figure 5.1: Mid-infrared spectra and machine learning analysis for classifying *An. funestus* mosquitoes based on infectious status. **A**, averaged mid-infrared spectra for infectious and non-infectious mosquitoes, which when analysed by the different machine learning algorithms, can enable categorisation of the mosquitoes based on their infectious status. **B**, accuracy of standard machine learning algorithms; K-Nearest Neighbours (KNN), Logistic regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), Random Forest (RF), and Multilayer perception (MLP) in distinguishing between infectious and non-infectious mosquitoes.

Our first XGBoost model, trained using the ELISA dataset, was able to predict the results of the ELISA test dataset with an overall accuracy of 92%. It classified spectra from the infectious and non-infectious mosquito samples with accuracies of 93% and 91% respectively (Fig. 5.2A). The same model was further tested to determine if mosquito age affected the classification, by introducing spectra from the known uninfected lab-reared 14-days old *An. funestus* from the laboratory (Table 5.2). The results showed that the performance was unaffected and was the same for classifying the new ELISA test dataset, suggesting that mosquito age did not confound the infection status in this model (Fig. 5.2B). The XGBoost model trained on ELISA data was also used to predict the infection labels of the spectra from mosquitoes screened for *Plasmodium* infection using PCR (Table 5.2). Here, the overall classification accuracy achieved by the model was 73% (Fig. 5.2C), though the model misclassified 43% of *Plasmodium*-negative samples (Fig. 5.2C); indicating limited generalisability of the model trained with ELISA derived data.

To understand the biochemical signature associated to this XGBoost model, we analysed the relative importance of specific spectral features highlighted by the model. We found that the X-H region of the MIR spectra (fundamental vibrations generally due to O-H, C-H, and N-H stretches) and fingerprint region (1500 - 500cm^{-1} frequencies) contributed most to the predictions (Fig. 5.3A).

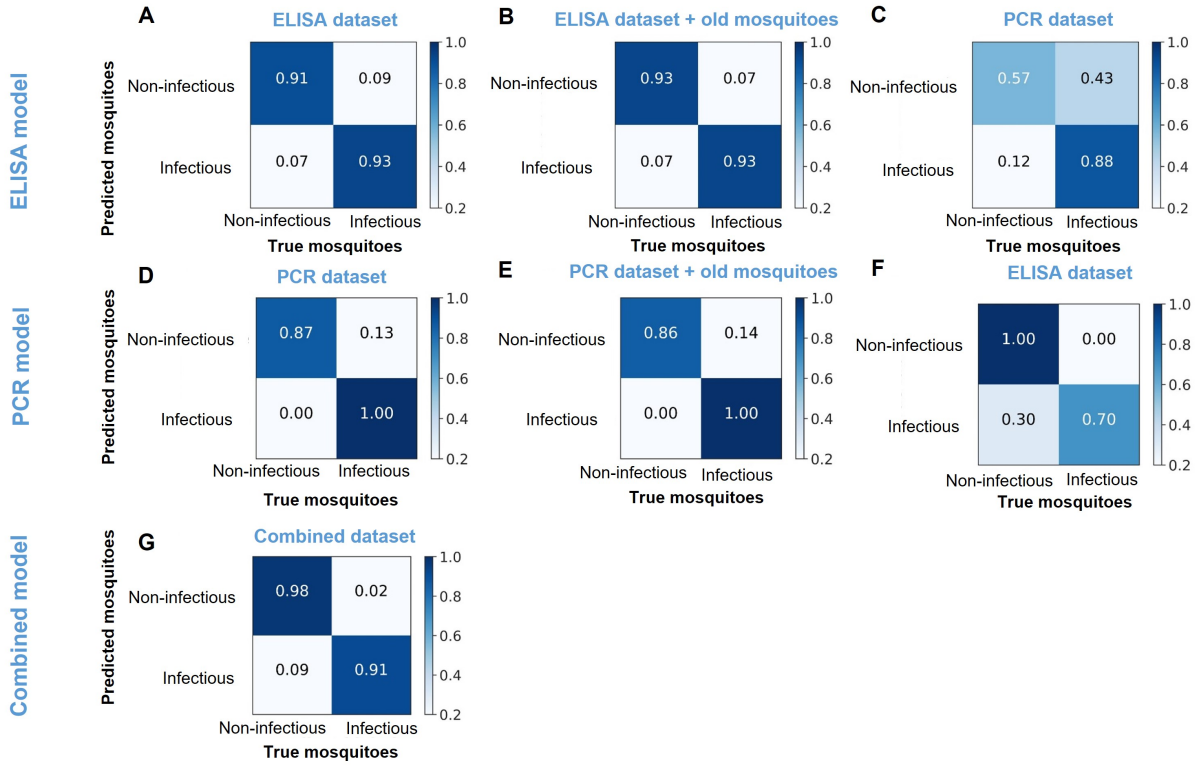

Figure 5.2: Illustrates the confusion matrices generated by the XGBoost model trained on ELISA and PCR infection datasets for predicting sporozoite infection in *An. funestus*. **A**, shows prediction results on an unseen segment of the ELISA dataset. **B**, displays predictions on augmented ELISA unseen dataset, including lab-reared 14-days old non-infectious mosquitoes. **C**, presents predictions on PCR dataset using the model trained on ELISA infection dataset. **D**, demonstrates predictions on the unseen segment of the PCR test dataset. **E**, shows predictions on a modified test dataset that integrates the PCR unseen test dataset with lab-reared 14-days old non-infectious mosquitoes data in the negative class. **F**, displays predictions on unseen ELISA test dataset using the model trained on PCR infection dataset. **G**, demonstrates the predictions from the model trained on the combined ELISA and PCR infection dataset for predicting sporozoite infection in *An. funestus*.

Our second XGBoost model, trained using the PCR dataset achieved an overall classification accuracy of 94% on the PCR test dataset, predicting infectious and non-infectious mosquito samples with 87% and 100% accuracies respectively (Fig. 5.2D). As above, to test the influence of mosquito age on the prediction, we incorporated some old non-infectious mosquitoes (i.e. age ≥14 days old) into a negative class to modify the PCR test dataset and found that the classification accuracy for this augmented test dataset was identical to the model trained without the augmentation (Fig. 5.2E). Finally, we tested this PCR-trained model for classifying the infectious and non-infectious samples in the ELISA-derived dataset, and found an 85% classification accuracy, with the model predicting infectious and non-infectious classes at 100% and 70% accuracies respectively (Fig. 5.2F). The results suggest that the model, compared to the ELISA-trained model, was more effective in differentiating between *Plasmodium*-negative and positive mosquitoes. This indicates its potential as a versatile tool for analysing samples screened with various

molecular techniques, including ELISA (see Fig. 5.2F). The analysis of important spectral features in this model showed that the spectral wavenumbers from ~2,000 cm^{-1} to ~700 cm^{-1} frequencies, which contain a complex series of absorptions, played a significant role in the predictions made by the XGBoost model (Fig. 5.3B).

To enhance generalisability, a new XGBoost model was trained using a combined ELISA and PCR dataset. This resulted in a prediction accuracy of 95% for the test data, including 98% accuracy for non-infectious mosquitoes and 91% for infectious ones (Fig. 5.2G). Notably, the crucial features contributing to this prediction, particularly from the X-H (encompassing O-H, C-H, and N-H stretching) and fingerprint regions, were also the key factors influencing the model predictions in the independent PCR and ELISA dataset trainings (Fig. 5.3C).

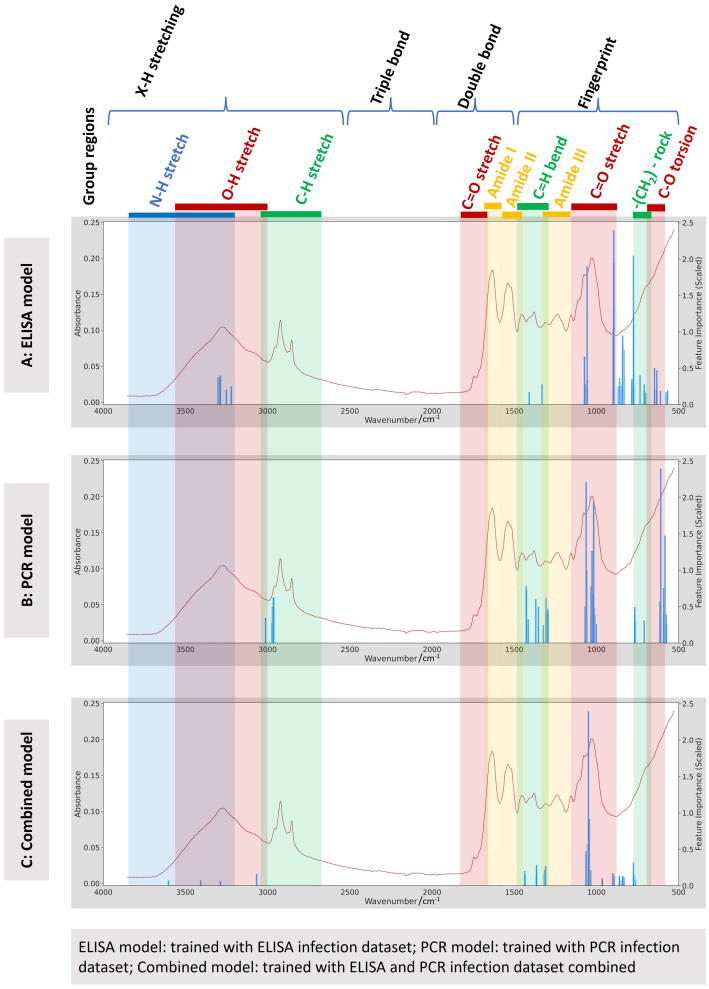

Figure 5.3: Illustrates the feature importance of the XGBoost model. The blue bars highlight the most important features for predictions, represented by scores assigned to each feature (wavenumber). The coloured stripes indicate the regions associated with different biochemical properties across the spectra. While the individual features may not be important on their own, their integration in the XGBoost Model enable the distinction of mosquitoes as either infectious or non-infectious.

5.3.3 Estimation of the entomological innoculation rate (EIR) from the balanced test sets of polymerase chain reaction (PCR) and enzyme-linked immunosorbent assay (ELISA) infection datasets

Estimation of the EIR was performed using balanced test sets from PCR and ELISA infection datasets used during model testing. Two parameters were used: sporozoite rate and biting rate. The sporozoite rate for PCR and ELISA was calculated as the number of infectious mosquitoes divided by the total number of mosquitoes tested (refer to Table 5.1). For MIRS prediction, the sporozoite rate was calculated as the number of mosquitoes predicted as infectious (sum of True Positives (TP) and False Positives (FP)) divided by the total number of mosquitoes predicted, as derived from the confusion matrices in Table 5.1. The low and high biting rates of 0.5 and 4.13, respectively, were sourced from literature as the biting rates for An. funestus in the Kilombero valley [128,218]. It was found that, in scenarios with both low and high biting rate, EIR estimates from the MIRS-ML models closely matched the 'ground truth' values from PCR and ELISA, showing minimal variation (Fig. 5.4).

Table 5.1: Displays the balanced, unseen segment of the PCR and ELISA infection datasets alongside their respective machine learning predictions

PCR model prediction on an unseen segment of the PCR infection dataset		
	Predicted non-infectious	**Predicted Infectious**
Actual non-infectious	TN = 13 (87%)	FP = 2 (13%)
Actual infectious	FN = 0 (0%)	TP = 14 (100%)
The total number of samples in the test set: N = 29 (Infectious = 14, Non-infectious = 15)		
ELISA model prediction on an unseen segment of the ELISA infection dataset		
	Predicted non-infectious	**Predicted Infectious**
Actual non-infectious	TN = 20 (91%)	FP = 2 (9%)
Actual infectious	FN = 1 (7%)	TP = 14 (93%)
The total number of samples in the test set: N = 37 (Infectious = 15, Non-infectious = 22)		
TN: True Negative, FN: False Negative, FP: False Positive, TP: True Positive		

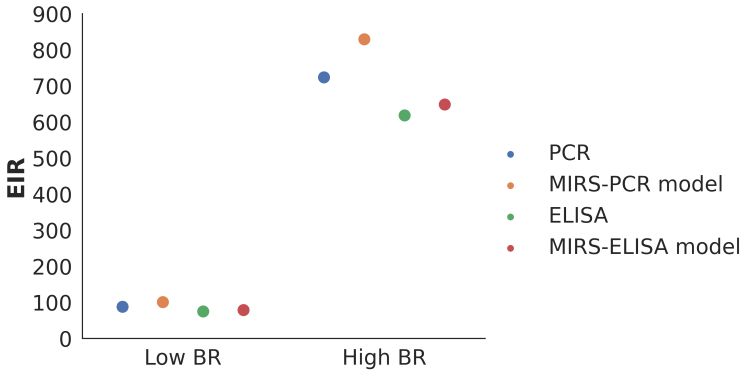

Figure 5.4: Estimated entomological inoculation rate from MIRS-ML, PCR, and ELISA predictions under hypothetical low and high mosquito biting rates.

5.4 Discussion

In the quest for effective malaria control, particularly in regions like sub-Saharan Africa where the burden of this disease is heaviest, the development of rapid, cost-efficient tools for monitoring transmission dynamics is imperative and urgent. Being able to swiftly identify *Plasmodium*-infectious *Anopheles* is particularly critical for understanding the transmission patterns in different localities, estimating the impact of interventions and planning new interventions. Unfortunately, current methodologies, predominantly ELISA and PCR, for detecting *Plasmodium* in *Anopheles* mosquitoes are resource-intensive, necessitating specialised skills and materials often scarce in local settings. This limitation hampers granular, district-level evaluations of malaria risk and the effectiveness of interventions.

Our research presents a novel, economical approach that leverages mid-infrared (MIR) spectroscopy coupled with supervised machine learning algorithms to swiftly identify *Plasmodium*-infectious *Anopheles* mosquitoes. By collecting and analysing the MIR spectral signatures from the heads and thoraces of wild-caught *An. funestus* females in rural Tanzanian villages, and subsequently validating these findings with ELISA or PCR for the presence of *P. falciparum* sporozoite, we established a reliable 'ground truth' for model training. The findings of this study are compelling, demonstrating that MIR spectral analysis can differentiate between infectious and non-infectious mosquitoes with accuracies exceeding 90% in certain cases. Notably, models trained on PCR data showed greater generalisability compared to those based on ELISA data, with mosquito age posing no significant interference. Although tested exclusively on *P. falciparum* and *An. funestus*, this advancement represents a significant step in malaria surveillance. Once calibrated for other major Afro-tropical malaria vectors and malaria transmission systems, it could have the potential to offer a scalable, low-cost solution that could transform data-driven decision-making in disease control programs. Moreover, we view this as an important step towards

creating a deployment-ready system but recognise that further development is necessary. Models trained using more diverse data from different settings will improve observed accuracies and enhance the readiness of this approach for broader implementation.

This study contributes to the expanding body of knowledge showcasing the potential of MIRS-ML based approaches for malaria vector surveillance. The use of these methodologies in delineating key entomological parameters such as age, species identification, and blood-feeding patterns of mosquitoes has been well documented [59,61,114]. The outcomes of our study suggest that this technology could serve as a versatile platform, enabling the interpretation of infrared scans to ascertain not only the species and age of mosquitoes, factors critical to their potential as malaria vectors, but also their blood-feeding history on humans or other vertebrates, and their infection status with malaria parasites. Such comprehensive profiling is instrumental in accurately characterising malaria risk, marking a significant advancement in vector surveillance and malaria control strategies.

In addition to the high classification accuracies of the MIRS-ML approach, the PCR-trained models also achieved generalisability of >85% in predicting sporozoite infection in wild-caught *An. funestus* mosquitoes even when predicting results of an ELISA dataset. These findings achieve consistent performance with studies utilising NIRS frequencies in the laboratory, which reported a >90% classification accuracy in detecting *P. falciparum* sporozoite infection in *An. gambiae* mosquitoes [93], and 77% accuracy in detecting *P. berghei* sporozoite infection in *An. stephensi* [215]. While earlier models trained on NIRS failed to identify mosquitoes infected with wild-strain parasites from asymptomatic malaria carriers, possibly due to limitations in the training dataset or detection capabilities of the system [219], models trained on MIRS, which provide clearer peaks with richer biochemical information appear to perform better [98,174]. This enhancement enabled our models to effectively identify infections in mosquitoes, a capability not fully realised with NIRS models in previous studies.

MIRS captures the biochemical composition of mosquito, which may consistently differ, in this case, with the infection status such as presence or absence of the parasite. The presence of parasite-specific proteins, such as circumsporozoite (CS) protein and the thrombospondin-related adhesive protein (TRAP), may contribute to the main spectral difference between infectious and non-infectious *An. funestus* [220]. Furthermore, since mosquitoes elicit immune responses to the parasites, this could consequently affect the biochemical characteristics of the infectious or non-infectious mosquitoes [221]. Additionally, higher levels of energy resource storage, such as glucose and lipid accumulation in the non-infectious mosquitoes [222,223], might yield distinct spectra signals between infectious and non-infectious *An. funestus*. This aligns with our observation where the majority of spectral features influencing machine learning prediction primarily originated from the O-H, C-H, and N-H bonds, as well as the

fingerprint region of the spectrum (1,500 cm^{-1} to 520 cm^{-1}), suggesting the presence of carbohydrates, protein, and lipids related to the parasite [115]. However, it is important to note that we are not focusing on individual spectra features; rather, we are using ML models to integrate a set of spectral features from different biochemical group regions to enable these classifications. While it may not be essential to identify specific features, we believe that additional studies should be conducted to better understand the biochemical signals underlying our algorithmic classifications.

The biological prerequisite that mosquitoes must exceed a certain age threshold (e.g. over 9 days) to become vectors for malaria transmission, due to the requisite extrinsic incubation period for the parasite [20], introduces potential age-related biases in detection efficacy. In this context, mosquito age could be considered a confounding factor influencing prediction accuracy. However, despite the theoretical possibility of age influencing the accuracy of predictions, our analysis demonstrated that the machine learning models adeptly identified signals indicative of infection across all age brackets, including older mosquitoes beyond 14 days, thus negating age as a significant confounding variable in our study.

Moreover, the ML model trained with the PCR infection dataset demonstrated an ability to generalise its prediction to samples screened by ELISA. In contrast, the model trained with the ELISA infection dataset had some limitation in predicting samples screened by PCR. We further observed similarities in the fingerprint region where both ELISA and PCR models detected signals, demonstrating agreement in parasite detection between the two models (Fig. 5.3). However, a noticeable difference was observed in the signals identified by the ELISA and PCR models, particularly in the frequency range of 3,500 to 3,000 cm^{-1}. Moreover, it is still not clear why ML models are picking up different signals from this region. Additionally, the generalisability of the ML model trained with PCR infection dataset can be attributed to the sensitivity of PCR in detecting even low sporozoite numbers in mosquitoes [224]. Leveraging the sensitivity of PCR can enhance the performance of MIRS-ML models. However, a study by Hendershot *et al.*, observed infection in mosquitoes at 0.5-1 day post-infection, indicating that false positive results can occur because PCR can report positives even when sporogonic development has not started [224]. This situation arises when an infectious blood meal has not full migrated to the mosquito abdomen, and the presence of gametocytes in the mosquito head and thorax is more likely to contribute to positive results.

The primary focus of this investigation was to showcase the field application of the MIRS-ML technique for detecting sporozoites in malaria vectors, not to directly compare it with PCR or ELISA methods, which were instead used solely to provide reference labels for ML model training. Moreover, this study represents only the first demonstration of field applications of the MIRS-ML technique for sporozoite detection in malaria vectors,

underscoring the need for further validation before its integration into surveillance or national malaria control efforts. Our analysis was also confined to *An. funestus* mosquitoes, chosen for their relatively high sporozoite rates in the region, highlighting the necessity to broaden future models to include more vector species. We recognise that expanding the MIRS-ML approach to all important mosquito species may necessitate compiling a comprehensive dataset of mosquito infection spectra, a task that presents logistical challenges in field settings, especially where natural infection prevalences are very low. A promising solution is to employ transfer learning, integrating laboratory-generated data with field-collected samples to enhance model robustness [61,178]. This method involves refining a model initially trained on laboratory data with new field data, facilitating the development of an effective tool for field infection prediction. Additionally, in low transmission setting where sporozoite infection rates are low, ELISA and PCR can be used for mosquito pool testing, reducing operational costs compared to individual mosquito tests. However, the feasibility of using MIRS-ML for mosquito pool testing remains unknown, prompting our investigation in the next steps. Additionally, in this study, MIRS-ML was not evaluated for identifying the *Plasmodium* species. Future studies should address this aspect to enhance the utility in regions where more than once species of *Plasmodium* is prevalent.

MIRS-ML proves cost-effective as it eliminates the need for repeated reagent costs in mosquito sample tests, with the only incurred expense being the initial £25,000/= for purchasing the FT-IR spectrometer. Capable of processing approximately 60 mosquitoes per hour, the portable bench-top design of the FT-IR spectrometer measures 22 x 33 x 26cm, requiring connection to an AC power supply. Currently, we are developing an online system to serve as a centralised platform for predicting various entomological and parasitological indicators of malaria. This online system aims to facilitate the scaling up of MIRS-ML, enabling end-users from different locations to upload unknown mosquito spectra for predictions related to infection status, species, age, or resistance status.

In conclusion, here we demonstrate the first application of mid-infrared spectroscopy combined with machine learning (MIRS-ML) for the rapid and accurate detection of *P. falciparum* in field collected *Anopheles funestus* mosquitoes. By analysing 7,178 female *An. funestus* specimens collected from rural Tanzania, we achieved detection accuracies of 92% and 93% against ELISA and PCR benchmarks, respectively. Moreover, MIRS-ML can guide programmatic decisions on vector control, as the EIR estimates, derived from MIRS-ML models, closely align with those obtained from PCR and ELISA methods across low and high biting rate scenarios, demonstrating consistency and reliability in malaria infection prediction. This method, capable of processing approximately 60-100 mosquitoes per hour with minimal costs, presents a significant advancement in malaria surveillance, particularly in sub-Saharan Africa where the disease has a profound impact. The utility of MIRS-ML extends beyond sporozoite detection, offering insights into critical entomological

indicators such as mosquito age, blood-feeding patterns, and species identification, thereby positioning MIRS-ML as a versatile tool in malaria risk assessment and evaluation of vector control interventions.

5.5 Methods

5.5.1 Mosquito collection and processing

Mosquitoes were collected from five villages in two rural districts in South-eastern Tanzania, Kilombero and Ulanga: Kisawasawa (7.8941°S, 36.8748°E), Mbingu (8.1952°S, 36.2587°E), Ikwambi (7.9824°S, 36.8216°E), Sululu (7.9973°S, 36.8317°E) and Tulizamoyo (8.3544°S, 36.7054°E) (Fig. 5.5). These villages experience annual rainfall of 1200 - 1800 mm, with mean daily temperatures of 20-32°C [225], and were selected because of the high densities of the malaria vector, *An. funestus*. The vector species was chosen for this study because it is the primary contributor to malaria transmission in the area and typically exhibits a higher prevalence of *Plasmodium* sporozoites compared to other local vector species such as *Anopheles arabiensis* [128, 130]. Mosquitoes were collected both indoors using CDC light traps and Prokopack aspirators [226, 227], and outdoors, in outdoor kitchens and animal sheds using resting buckets [228]. The collected *Anopheles* mosquitoes were sorted by taxa based on their morphological features [202]. All *An. funestus* group mosquitoes were immediately killed with chloroform and then stored individually in 1.5mL microcentrifuge tubes containing silica gel as desiccant and preservative. The *An. funestus* samples were transferred to the VectorSphere laboratory at the Ifakara Health Institute and stored dry for at least five days for further investigation.

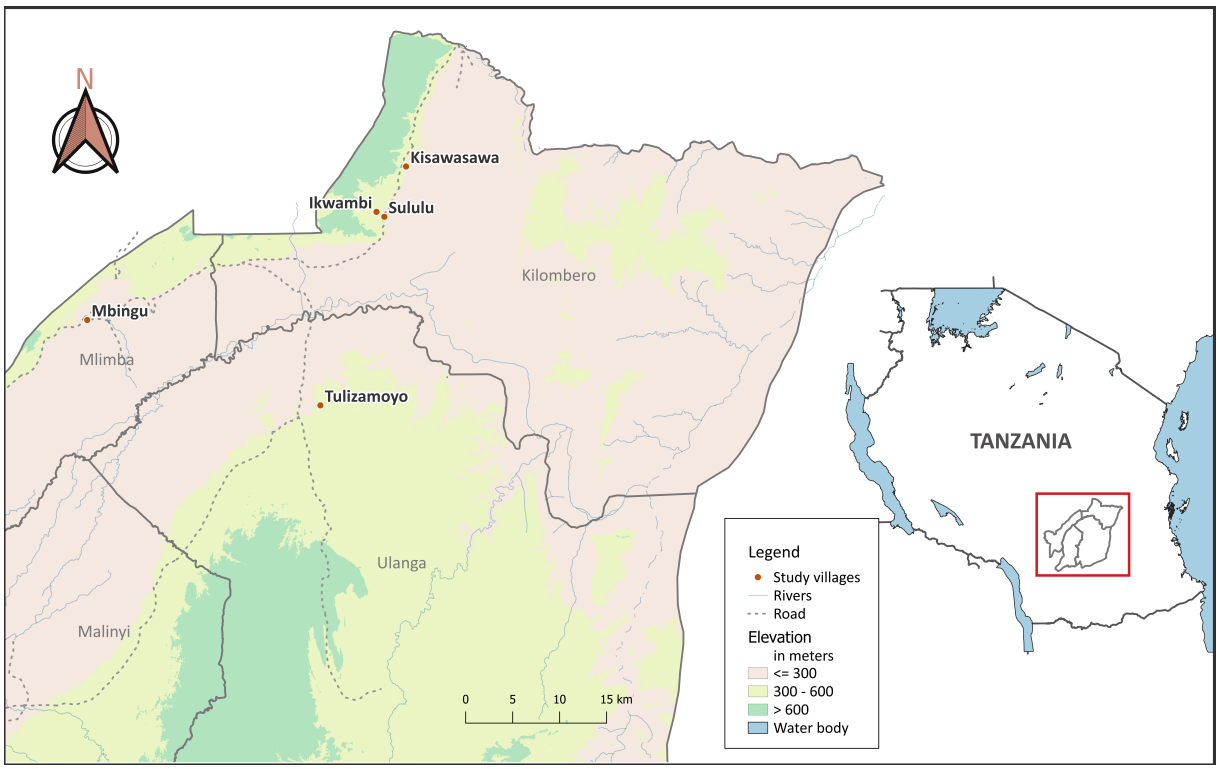

Figure 5.5: Map of the five villages where mosquitoes were collected.

5.5.2 Mid-infrared spectroscopy

We used a Bruker ALPHA II Fourier-Transform Infrared (FT-IR) spectrometer with attenuated total reflectance (ATR) to measure the infrared spectrum of the dried mosquito samples. Prior to scanning, the head and thorax of each mosquito was carefully separated from the abdomen, ensuring that only the head and thorax regions were scanned. The mosquito heads and thoraces were placed on the infrared optical window, and pressure was applied to ensure maximum direct contact between the sample and the diamond crystal. The spectral signal was obtained at frequencies between 4,000 to 400 cm^{-1}, with a resolution of 2 cm^{-1}. Each spectrum was an average of 32 scans of a single mosquito sample, with band intensity recorded as an absorbance. Following scanning, the remaining of the mosquito head and thorax (carcasses) were individually packed in 1.5mL tubes for subsequent molecular analysis. The recorded spectra were pre-processed by compensating for carbon dioxide interference bands and water vapour absorption bands as previously described [59]. Additionally, spectra with no intensity (i.e. flat spectra) and low intensity (i.e. <0.11 absorbance units) were removed before machine learning steps [59].

5.5.3 Detection of Plasmodium sporozoites using polymerase chain reaction (PCR) and enzyme-linked immunosorbent assay (ELISA)

To obtain reference labels of *P. falciparum* sporozoite infections in the mosquito head and thorax carcasses, we used real-time PCR targeting var gene acidic terminal sequences (varATS) of the parasites [78,214] and Enzyme-linked immunosorbent assays (ELISA) assays for detecting circumsporozoite protein (CSP) [75]. Each carcass underwent individual analysis, a method previously demonstrated for detecting mosquito blood meal remaining [229]. In total, 7,178 *An. funestus* carcasses were examined across two rounds using the following methods:

Initially, 4,281 samples were screened using ELISA [75], each time ensuring that the lysates of all the positive samples were boiled for 10 minutes at 100°C to eliminate false positives usually associated with heat labile non-*Plasmodium* protozoans, and retested [85].

Next, 2,897 samples underwent multiplex real-time PCR for sporozoite infection detection. This involved DNA extraction of mosquito carcasses using the DNAzol® reagent [230], DNA was eluted in 50μL of Tris-Acetate-EDTA (TAE) buffer. Subsequent to this, Real-Time PCR was conducted targeting the Pan-*Plasmodium* 18S rRNA and *P. falciparum* specific *varATS* sequences, along with a 28S rRNA mosquito sequence as a reference gene/internal control, enhancing specificity and sensitivity for *P. falciparum* and non-*falciparum* species.

The PCR reaction used a 10μL mix including Luna Universal Probe qPCR Master Mix (New England Biolabs, USA), a primer mix, water, and template DNA. The thermal cycle parameters involved an initial polymerase activation at 95°C for 1 minute, DNA denaturation at 95°C for 15 seconds for 45 cycles, and annealing/elongation at 57°C for 45 seconds for 45 cycles [78,231]. Samples exhibiting a sigmoid curve that reached the cycle threshold (Ct) value at $\leq$35 cycles were classified as positive, while those reaching >35 cycles were classified as negative. The assays were run in duplicates, and each run included a non-template control and *P. falciparum* NF54 DNA as positive control. The real-time PCR measurements were analysed using CFX96 Real-Time PCR system (Bio-Rad Laboratories, USA).

5.6 Data analysis

The PCR and ELISA data were separately used as references to evaluate performance of the infrared spectroscopy and machine leaning models for accurately identifying individual mosquitoes infected with *P. falciparum* sporozoites in their salivary glands. Since only a small proportion of the mosquitoes were found infectious (see results section), it was necessary to first obtain similar numbers of randomly selected non-infectious mosquitoes as controls, to avoid skewed model performance. The non-infectious samples were therefore under-sampled by randomly selecting individual specimen based on their smallest average Euclidian distances to the 3 farthest positive samples [203, 232]. This process was repeated 50 times and bootstrapped to cover as many negative samples as possible.

To ensure consistency and uniformity, the spectra data were standardised using the StandardScaler algorithm [141]. Supervised machine learning techniques, including K-nearest neighbours (KNN), logistic regression (LR), support vector machine (SVM), gradient boosting (XGB), random forest (RF), and multilayer perception (MLP), were then compared for predicting the ELISA and PCR results. The model with the highest accuracy was optimised further by adjusting its hyper-parameters using randomised search cross-validation, and its final estimator was evaluated using K-fold cross-validation (k = 5). The analysis was performed using Python 3.8 with the Scikit-learn library [141]. The machine learning models were trained using ELISA, PCR, or combined ELISA + PCR training datasets and tested on all three corresponding test sets (Table 5.2). Training was done with up to 90% of the known positive and negative samples, each time leaving out at least 10% for model validation (Table 5.2). Additional validation of the models included using samples tested by either of the two methods, and incorporating lab-reared, non-infectious mosquitoes confirmed to be at least 14 days old. This was to guarantee that the models accurately classified infection status rather than mosquito age, as age can confound results and only mosquitoes older than 9 days are capable of transmitting malaria [20].

Table 5.2: Training and test datasets used in the different models

Model	Training data	Test data
ELISA	ELISA dataset (90%)	1. ELISA dataset (10%) 2. ELISA dataset (10%) modified (14-day oldlab-reared *An. funestus* mixed into negative class) 3. PCR dataset
PCR	PCR dataset (90%)	1. PCR dataset (10%) 2. PCR dataset (10%) modified (14-day old lab-reared *An. funestus* mixed into negative class) 3. ELISA dataset
Combined	PCR & ELISA dataset combined (90%)	1. PCR & ELISA dataset combined (10%)

5.7 Ethics approval and consent to participate

Ethical approval for this study was obtained from the Institutional Review Board at Ifakara Health Institute (Ref. IHI/IRB/No: 41-2020), and the Medical Research Coordinating Committee (MRCC) at the National Institute of Medical Research (NIMR) (Ref: NIMR/HQ/R.8a/Vol. IX/3557). Since this study focused primarily on malaria mosquitoes, it did not involve human participants or animals.

5.8 Code, data, and materials availability

The mid-infrared spectral datasets generated and analysed during the current study, as well as code for the analyses is available at https://github.com/MwangaEP/Sporozoite-detection-funestus.

5.9 Competing interests

The authors declare that they have no competing interests

5.10 Funding

This study was supported by the Wellcome Trust Masters Fellowship in Tropical Medicine & Hygiene (Grant No. 214643/Z/18/Z) awarded to EPM, and the Medical Research Council (MRC) [MR/P025501/1]. FB was supported by the Academy Medical Sciences Springboard Award (ref:SBF007/100094). FOO was supported by a Howard Hughes Medical Institute (HHMI)-Gates International Research Scholarship (Grant No. OPP1099295), and Bill and Melinda Gates foundation (INV003079). SAB was supported by the Bill and Melinda Gates Foundation (INV-030025) and Royal Society (ICA/R1/191238).

5.11 Authors' contributions

EPM, SB, FB, and FOO conceived the study. EPM, SA, FOO, FB and DJS developed the study's protocol. EPM, SA, FEM, ISM, and PAK and performed molecular assays. EPM collected the data and carried out data analysis and ML training. EPM wrote the manuscript. EPM, GS, SHM, IHM, MGJ, KW, SB, FB and FOO reviewed and edited drafts of the manuscript. All authors have read and approved the final manuscript.

Chapter 6

Lessons Learned and Future Prospects for Mid-Infrared Spectroscopy in Malaria Surveillance

Emmanuel P. Mwanga[1,2*], Issa H. Mshani[1,2], Simon A. Babayan[2], Francesco Baldini[1,2], Fredros O. Okumu[1,2,3,4]

1. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, Morogoro, Tanzania.
2. School of Biodiversity, One Health and veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, UK.
3. School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
4. School of Life Science and Bioengineering, The Nelson Mandela African Institution of Science and Technology, P. O. Box 447, Arusha, Tanzania.

6.1 Abstract

Effective surveillance is crucial for malaria elimination with key metrics including parasitological (e.g., parasite incidence and prevalence in humans) and entomological (e.g., prevalence of *Plasmodium* sporozoite infections in *Anopheles* mosquitoes, mosquito age and human blood index) factors. Traditional methods for gathering these data include PCR, microscopy, rapid diagnostic tests (RDTs), and ELISA for parasitological estimates, and dissections, microscopy, ELISA, and PCR for entomological estimates. Though effective, these methods are labour-intensive, slow and costly. Other challenges include inadequate expertise, need for frequent resupply of reagents, and poor sensitivity and specificity e.g., those caused by cross-reactivity in ELISA assays (false positives) or parasite mutations in some RDTs tests (false negatives). Infrared spectroscopy (IR), particularly when paired with chemometrics or machine learning (ML), offers a rapid, low-cost, and reagent-free alternative for measuring multiple malaria indicators associated with chemical changes in the samples. In particular, techniques using near-infrared or mid-infrared spectra have been used to determine mosquito age, species, infection status,

and blood meal sources. This chapter discusses the lessons learned from the different research studies using IR spectroscopy; and broadly explores the future prospects for IR-spectroscopy and machine learning for malaria vector surveillance. While recognising significant recent advances, there are still several challenges that must be addressed to ensure optimal performance – notably improved model generalisability for different use cases and interpretability of bio-chemical signals captured by the infrared spectra. Techniques such as transfer learning can enhance model performance across different environments but there are still no effective approaches that can fully address the broader variability of field samples. Broader implementation of MIRS-ML for malaria surveillance will require continuous, extensive data generation and model validation. To achieve the scale-up, future research should also focus on developing deployment-ready systems and inclusion of pooled samples, as current scanning is limited to individual samples. Lastly, while current ML models have not achieved satisfactory diagnostic accuracy levels and are therefore more useful for field screening than diagnostics, the approaches hold potential as a surveillance tool.

6.2 Background

Achieving malaria elimination requires effective surveillance techniques to measure biological attributes that influence the overall potential of malaria transmission and to assess impacts of interventions. These attributes can be parasitological, e.g., parasite incidence and prevalence in humans, or entomological, e.g., prevalence of *Plasmodium* sporozoite infections in *Anopheles* mosquitoes, proportions of mosquito blood meals derived from humans other than other vertebrates (human blood index), and mosquito age [33, 136, 185, 233, 234]. For many years, polymerase chain reaction (PCR) and enzyme linked immunosorbent assays (ELISA) have been the cornerstone for vector and parasite surveillance, detecting parasite and host blood-meal in mosquitoes [68, 69, 75, 78, 79, 213], as well as parasite in human blood [214, 234]. Additionally, dissections are commonly used for assessing ovarian development stages to inform age classification and mating status.

Though broadly effective, these current approaches are labour-intensive and time-consuming, with high operational costs due to the costs of equipment installation and repeated need for reagents in routine surveillance. The development of malaria rapid diagnostic tests (RDTs) has revolutionised parasite detection in human blood due to their simplicity and accessibility [235–237] but no similar system currently exists for detecting *Plasmodium* infections in mosquitoes. Moreover, in areas with low transmission and the risk of Histidine-rich protein II (HRP-2) gene deletions, there is a potential for missed malaria infections, leading to significant false-negative results [236, 238–241]. The need for

effective surveillance therefore remains one of the key priorities for countries aiming at malaria elimination.

In 2015, the World Health Organisation, in its global technical strategy for malaria elimination, recommended that endemic countries prioritise transforming malaria surveillance into a key intervention in their elimination policies, supporting innovation and research to develop surveillance tools and strategies that capture essential malaria data in a complete, accurate and timely manner [32]. In response, countries must find scalable approaches to conduct such surveillance and effectively use the data to improve their program outcomes cost-effectively. Unfortunately, most malaria endemic countries do not yet have adequate surveillance systems and are not measuring all the essential malaria metrics consistently [34]. The authors collected country metadata on vector surveillance and control activities, via an online survey by officials from National Malaria Control Programmes (NMCPs) and partner organisations and analysed these activities for alignment with WHO-recommended indicators. This study revealed significant differences between countries approaching malaria elimination and those with intense transmission; and identified gaps in data collection and management strategies [34]. In a separate analysis [242], it was found that most vector surveillance programs lack sufficient capacity, with only 80% of NMCPs having the necessary resources. Moreover, countries nearing malaria elimination tended to have more operational staff and better training systems than countries in malaria control phase [242]. The authors concluded that strategic planning and training deficiencies are major obstacles to effective vector surveillance [242]. Addressing these gaps broadly requires simpler low-cost approaches that can be deployed at scale and without requiring extensive staff training.

6.3 Applications of infrared spectroscopy and machine learning for entomological surveillance

In recent years, infrared spectroscopy (IR) has gained attention for its ability to measure key entomological and parasitological indicators of malaria transmission [204]. Unlike commonly used methods that require multiple reagents, IR, when combined with chemometrics or machine learning (ML), offers a quick, low-cost and reagent-free approach, requiring only desiccants [204]. The IR assesses the biochemical composition of biological samples (i.e., protein, lipids and carbohydrates) [59, 116], providing insights into various factors including the age and species of mosquitoes [59–61, 94, 147, 175–177, 243, 244], presence or absence of the pathogen infection in mosquitoes and humans [92, 93, 107, 113, 115, 126, 215, 245, 246], blood meal sources of field-collected mosquitoes [229], and the symbiont Wolbachia in mosquitoes [95, 247]. The two main types of IR spectroscopy employed to measure these

indicators of malaria transmission include Near-infrared spectroscopy (NIRS) and Mid-infrared spectroscopy (MIRS).

The NIRS approach, as used in entomological and parasitological studies, is a rapid, non-destructive technique that requires no sample preparation, not even desiccation. It operates within the near-infrared wavelength, ranging from 12,500 cm^{-1} to 4,000 cm^{-1} wavenumbers [98]. NIRS absorptions are mainly based on overtone and combination vibrations within the sample [98]. Overtone absorption bands are due to vibrational transitions from the ground state to higher excited states. They appear at frequencies that are approximately integer multiples of the fundamental vibrational frequency and are typically weaker in intensity [97,174]. Combination modes occur when different vibrations are excited simultaneously, resulting in combined vibrational modes [97,174]. The application of NIRS combined with multivariate statistics, partial least square regression (PLS) analysis has been demonstrated severally for determining species and age of major Afro-tropical malaria vectors, *An. gambiae s.s* and *An. arabiensis* [60,90,94,176,177]. Other studies have also shown the application of NIRS for detecting *P. falciparum* infections in mosquitoes and human blood [92,93,107,245], as well as for detecting Wolbachia, chikungunya and Zika virus in *Aedes aegypti* mosquitoes [246–248].

In more recent studies, further improvements in measuring the entomological and parasitological indicators of malaria transmission have been achieved by focusing on the corresponding fundamental vibrations in MIRS, which operates within a wavelength of 4,000 cm^{-1} to 400 cm^{-1} wavenumbers [97,98], and can address the challenge of the weaker nature of absorption overtones and combination bands. For example, Gonzalez-Jimenez *et al.*, demonstrated a machine-learning approach using mid-infrared spectra to simultaneously identify age and species of *An. gambiae* and *An. arabiensis* mosquitoes, achieving 82.6% species classification accuracy [59]. In a follow-up study, Siria *et al.*, used the convolutional neural network (CNN) to learn from MIRS spectra data and predict both age and species from single samples of *An. gambiae*, *An. arabiensis*, and *An. coluzzii* originating from laboratory and semi-field conditions with up to 95% prediction accuracy [61]. Another study has further demonstrated the feasibility and potential of MIRS-ML for age grading by extending its application to predict the epidemiologically relevant age categories (young and old) of another Afro-tropical malaria vector, *An. funestus* [244].

Additionally, ML models have been shown to distinguish host-blood meal sources in the abdomens of laboratory reared *An. arabiensis* with an accuracy of 98% [114]. Although MIRS-ML shows great promise as a surveillance tool, validations in field conditions are important. A recent study demonstrated a field application of MIRS-ML in detecting host blood-meal sources in field-collected *An. funestus*, achieving a classification accuracy of 90% [229]. The study also showed that the HBI estimates derived MIRS-ML models and PCR estimates are closely aligned, suggesting consistency between the MIRS-ML

models and the common PCR approach. Furthermore, MIRS-ML was also effective in detecting sporozoite infections in naturally infected *An. funestus* [249]. More research has also focused on parasite detection in human-blood. For example, MIRS combined with PLS was used to detect early ring stages of *P. falciparum* cultured in the laboratory, achieving detection limits of less than 1 parasite/microliter(μL) [113]. In a subsequent study, *Plasmodium* cultures were introduced into whole blood samples from uninfected individuals, and by analysing MIR spectra data with PLS, they accurately detected 98% of specimen with parasitaemia densities above 0.5% [126]. However, these studies were limited by their reliance on laboratory cultures, which may not capture the genetic diversity present in natural environment. In contrast, Mwanga *et al.*, detected *P. falciparum* in human dried blood spots (DBS) obtained from a cross-sectional malaria survey in 12 wards in south-eastern Tanzania, achieving an overall accuracy of 92% in distinguishing malaria infected and uninfected DBS [115].

Despite the notable accuracies, one major challenge reported by researchers has been the lack of generalisability between sites [61]. The models often failed to accurately predict unseen data from different locations, such as different laboratories, and semi-field settings, likely due to the inherit variability in mosquito samples [61]. This variability potentially originates from differences in environmental conditions, such as temperature and humidity, which can impact mosquito development rates and physiology. Variation in mosquito populations, including species composition and genetic diversity, can also lead to differences in the biochemical composition of the sample. Furthermore, dietary factors, such as the availability of blood meals and sugar sources, can affect mosquito development, fitness and survival, altering the characteristics of the sample used for model training and testing. These factors can create discrepancies in the data, reducing ability of the model to generalise across different settings. As a result, models that performed in one setting may not necessarily translate to effective prediction in another, highlighting the need for more robust approaches to address these discrepancies. To improve generalisability of the ML models in predicting samples originating from different locations, transfer learning can be performed by updating pre-trained models using small subsets of data from the target population to improve prediction on unseen data in that target population [61,178].

To advance the utility of these approaches and fulfil the need of malaria endemic countries for scalable surveillance systems that are both effective and low-cost, it is crucial to evaluate key lessons from the use of infrared spectroscopy and machine learning for vector and parasite surveillance. Additionally, it is important to identify essential aspects for consideration to ensure future integration into malaria surveillance programs, including addressing protocols (such as sample preservation and scanning), instrument installations and maintenance, data management and processing, machine learning analysis, and transfer learning. Additionally, challenges such as the interpretability of

biological signals, the generalisability of MIRS-ML, and its implementation must be considered. The following sections will discuss these lessons and challenges in detail.

6.4 Key lessons from applications of infrared spectroscopy in malaria vector surveillance

Previous research on infrared spectroscopy has highlighted critical lessons in protocol development, instrument maintenance, and data analysis. Key insights include best practices for sample preservation and scanning, which ensure consistent and accurate measurements. For purposes of this thesis, the focus is solely on entomological applications of these technologies and does not include work on parasitological assessments in humans.

6.4.1 Preparation and preservation of mosquito samples

Mosquito collection methods vary based on the indicators being assessed. For evaluating biting rates, host-seeking mosquitoes are collected using human landing catches, odour-baited or human-baited traps, such as CDC light traps placed near occupied bed nets, or volunteer-occupied double net traps [166, 227].

For NIRS or MIRS studies aimed at accurately predicting the age, species, and blood-feeding histories of collected mosquitoes, the ideal method for killing should possess several key characteristics (Table 6.1). It should be inert, meaning it does not chemically alter or leave residue on the mosquito samples. The method should be safe for researchers to handle, non-toxic, and should not pose significant health risks. Importantly, it should have minimal impact on the mosquito's biochemical composition, ensuring that it does not interfere with infrared spectra. Finally, the method should kill the mosquito while also preserving the samples against decomposition, thereby maintaining the integrity of the data for accurate predictions.

Currently, chloroform is commonly used in entomological studies for killing mosquitoes [61, 114, 250–252]. Chloroform is favoured because it is fast, efficient, and does not leave residues on the mosquito samples. Additionally, it kills bacteria within the mosquito, minimising decomposition and thereby having minimal impact on MIRS or NIRS prediction. However, chloroform is also known to be toxic and carcinogenic upon inhalation or prolonged exposure [253], posing health risks to researchers.

An alternative approach is the use of insecticide formulations, such as Kaltox Paalga® (which includes active ingredients pyrethroids, allethrin 0.27%, permethrin 0.17%, tetramethrin 0.20% and an organophosphorus compound, chlorpyrifos ethyl 0.75%). A study comparing this insecticide to chloroform for killing mosquitoes before NIRS scanning found no significant difference in NIRS prediction accuracy, with chloroform and the insecticide yielding classification accuracies of 92% and 90%, respectively [254]. However, the increasing resistance of mosquitoes to insecticides necessitate the exploration of non-pyrethroid formulations.

Other methods, such as ethanol or freezing at -20°C, are also commonly used. Ethanol, however, is not ideal for infrared spectroscopy as it denatures proteins and washes away lipids in biological samples [255]. Freezing has been evaluated primarily for preserving mosquitoes after killing rather as a method for killing mosquitoes [256]. The impacts of these alternative mosquito killing methods on NIRS or MIRS predictions require further investigation.

Table 6.1: Comparison of mosquito killing methods for NIRS/MIRS studies

Method	**Chemical alteration**	**Human safety**	**Effect on IR spectra**	**Notes**
Chloroform	No	No	None	Toxic; carcinogenic; handle with care.
Insecticide	No	Yes	None	Depends on formulations; explore non-pyrethroids.
Ethanol	Yes	Yes	High	Denatures proteins; not ideal for MIRS studies.
Freezing (-20°C)	No	Yes	Minimal	Introduce moisture; noise in MIRS spectra.

After mosquitoes are killed, the ideal preservative for infrared studies should preserve the biochemical integrity of the samples, ensuring that the chemical composition remains unaltered for accurate prediction of entomological indicators of malaria transmission. Various preservatives have been used in previous studies including RNAlater, ethanol, Carnoy's fixative (3:1 ethanol: acetic acid), silica gel (silicon dioxide (SiO_2)), anhydrous calcium sulfate, refrigeration at 4-5°C, and freezing at -20°C [250–252, 254, 256]. Most studies on NIRS indicate that the best results for age classification and species identification in mosquitoes are achieved when samples are preserved using silica gel, refrigeration, or RNAlater [250–252].

Silica gel is porous and highly absorbent, making it effective for keeping products dry and free from mould (Fig. 6.1A). Its inert nature ensures that it does not chemically interact with mosquito tissues, preserving the integrity of the samples. Silica gel is also low-cost, widely available as commonly used in packaging and laboratories, and suitable for long term preservation. Mgaya *et al.*, demonstrated that silica gel can preserve mosquito samples

up to 8 weeks without introducing noise to the spectra, making it ideal preservative for NIRS and MIRS studies [256].

Freezing at -20°C is another method used for preserving mosquito samples. While it effectively halts biological process and prevents microbial growth, freezing can introduce excess water into the samples. This added moisture can cause noise in the MIRS spectra, affecting the accuracy of prediction [59]. Additionally, repeated freeze-thaw cycles can degrade the sample quality, making freezing less ideal for long-term preservation compared to silica gel.

RNAlater is effective at stabilising ribonucleic acid (RNA) and preventing degradation (Fig. 6.1B), which helps maintain the biochemical composition of the samples. However, its effectiveness in infrared-based studies decreases for storage duration exceeding four weeks [251], and it is less cost-effective compared to Silica gel and freezing [250, 251].

Ethanol, though commonly used for preserving biological samples (Fig. 6.1C), can be problematic for IR studies. Ethanol depletes lipids and denatures proteins [255], which are critical for accurate MIRS predictions, leading to alterations in the biochemical composition of the biological samples. Therefore, while ethanol is readily available and effective in killing mosquitoes, it is less suitable for preserving samples for infrared spectroscopy. Carnoy's fixative, is the mixture of ethanol and acetic acid, is effective for preserving nucleic acid and prevent autolysis. However, the acetic acid component as it is ethanol may cause degradation of lipids and proteins, potentially altering the infrared spectra. This also makes Carnoy's fixative less suitable for preserving samples intended for infrared studies, where maintaining the original biochemical composition is important.

The most comprehensive assessment of these preservatives for MIRS was done by Mgaya *et al.*, who evaluated the effect of sample preservation methods and storage duration on the performance of MIRS for predicting the age of malaria vectors, *An. arabiensis*. In their study, laboratory-reared mosquitoes were first killed using ethanol then preserved using either silica gel desiccant, freezing, or ethanol. The study found that silica gel consistently proven to be the most suitable preservative method. Additionally, the study found that the highest accuracy in age prediction was achieved when models were trained and tested on similarly preserved samples, but classification accuracy declined significantly when training and test samples were preserved differently [256]. This emphasises the need the need for standardised sample-handling protocols in infrared studies to ensure consistency and accuracy in predictions. Table 6.2 summarises the pros and cons of the preservatives used for mosquito sample storage in infrared studies.

Figure 6.1: Mosquito preservation methods: A) Silica gel, B) Ethanol, and C) RNAlater.

Based on these assessments, we recommend maximising the potential of NIRS and MIRS for predicting mosquito age, species, or blood-feeding histories by killing mosquitoes with chloroform, ensuring sufficient caution, or alternatively using ethanol or freezing. However, samples should be preserved using silica gel, which is widely available, effective, and does not alter the biochemical composition of the samples. Other approaches, such as RNAlater, ethanol or freezing, can be considered only under specific circumstances if required, but their limitations should be carefully considered based on goals of the study and duration.

Table 6.2: Preservatives used for mosquito sample storage in infrared studies.

Preservative	Characteristics	Pros	Cons
Silica gel	Desiccant	Inexpensive, effective for long-term storage	-
RNAlater	RNA stabilizer	Preserves biochemical composition	Decrease in effectiveness with time, costly.
Ethanol	Alcohol	Readily available	Lipid depletion, protein denaturation.
Carnoy's fixative	Alcohol and acetic acid	Preserves nucleic acids	Potential alteration of lipids and proteins.
Freezing (-20°C)	Temperature-based	Prevents microbial growth	Introduce moisture and noise in MIRS spectra, costly as requires electricity for refrigerators; may cause mould and fungus to the sample.

6.4.2 Infrared scanning of mosquito samples

Infrared scanning of mosquito samples allows for the measurement and collection of spectra from various body parts, including the head and thorax, abdomen, and legs. For instance, when identifying infectious mosquitoes, the head and thorax are also suitable

because mature *Plasmodium* sporozoites are typically lodged in the salivary glands within the mosquito mouth parts [92, 93, 249]. For identifying mosquito age and species, the head and thorax are ideal because pteridines concentration, which are linked to ageing, accumulate in these body parts due to the presence of thick tissues [60, 61, 244, 257]. However, other mosquito body parts like legs can also be used for age and species prediction, as demonstrated in various studies [258–260], as well as cuticular resistance which is linked to the thickening of legs [117, 261]. The abdomen is best for measuring mosquito blood-feeding histories as the gut content containing blood reveals differences between various host blood meal sources [114, 229]. In future such abdominal measures might be extended to inform not just the source of the blood meals but also the digestion stage of the blood meal.

The speed and efficiency of sampling are influenced by both the modality and the type of instruments used. NIRS is widely used due to its ability to quickly scan large numbers of samples with minimal preparations. Most studies employing NIRS utilise the QualitySpec Pro instrument, where up to 20 mosquitoes can be placed on a plate, with each mosquito scanned individually [60]. The NIR spectrometer with fibre optic probe collects a minimum of 20 spectra from each sample, which are then averaged into a single spectrum. This technique is highly efficient, enabling the scanning of ~100 freshly collected mosquitoes per hour, as long as they are immobilised. The speed and ease of use make NIRS ideal for large-scale studies, particularly when fresh samples are available.

For MIRS, attenuated total internal reflectance – Fourier transform infrared (ATR-FTIR) spectrometers are commonly used because they allow for fast and stable collection of MIR spectra. In this system, samples are placed directly on top of the ATR diamond crystal, and an anvil (an adjustable pressure arm) ensures maximum contact between the sample and the ATR crystal. This step is crucial because only the evanescent waves of infrared light, extending approximately 0.5 to 5 μm beyond the ATR crystal, penetrate the sample before being refracted back into the crystal [262]. This technique allows for more detailed chemical analysis, as MIRS wavelength provides information about the molecular composition of the sample. The spectrometer can scan 60-100 samples per hour, depending on factors such as the number of scans per sample and speed of the operator [249]. The precision of ATR-FTIR is one of its main advantages, as it allows for the detection of subtle changes in the chemical composition of the sample. However, it may require more careful sample handling and preparation to ensure consistent results.

While both NIRS and ATR-FTIR offer distinct advantages, the choice between them depends on the specific need of the study. NIRS is preferable for high-throughput applications where speed, ease of use, and non-destructiveness to the sample are paramount, while ATR-FTIR is more suitable for detailed chemical analysis where molecular information is critical. The limitation of each method, such as the potential for

surface-level analysis in NIRS or the need for precise sample placement in ATR-FTIR, should be considered when selecting the appropriate techniques for infrared studies. Table 6.3 summarises the basic characteristics of NIRS and MIRS (ATR-FTIR).

Table 6.3: Some characteristics of NIR and MIR (ATR-FTIR)

	MIRS (ATR-FTIR)	**NIRS**
wavenumber	400 - 4,000 cm^{-1}	4,000 - 12,500 cm^{-1}
Absorption bands due to	Absorbed radiation (fundamental vibration)	Absorbed radiation (overtones and combination)
Absorption bands	Well-resolved, assignable to specific chemical groups	Series of broad overlapping bands
Signal intensity	Good, More intense than NIRS	Good
Interference	Atmospheric intrusion such as water, carbon dioxide and humidity, physical attributes (e.g., sample size, shape, and hardness)	Water, physical attributes (e.g., sample size, shape, and hardness)
Sample preparation	Reduced, essential with ATR	None

In the MIRS systems, multiple scans are essential for improving the signal-to-ratio and ensuring the reliability of the spectra data. Each scan records the infrared absorption of the sample, and by averaging multiple scans, random noise can be minimised, leading to a more accurate and stable spectrum. Depending on the specific type of study and tissues being scanned, background and sample measurement can be taken with varying number of scans. For instance, 64 sample scans, which last for 1 minute, provide highly reliable data at the cost of longer processing time. Conversely, 32 sample scans last for 30 seconds, offering balance between speed and accuracy, while 16 sample scans, taking only 15 seconds, make the spectrometer rapid while maintaining sufficient efficiency for many applications [262], including mosquito studies [61]. The average spectrum from these scans is stored at a resolution of 2 cm^{-1}, which can be generally adequate for most analytical purposes.

Going forward, we recommend that every individual mosquito should be scanned in multiple body parts, including at least head and thorax, and the abdomen, and that the data can be used separately to answer different questions and provide measures different surveillance indicators. Such an approach would save the time for and enable the extension of this technology to be multi-purpose and more comprehensive. Additionally, the ideal direction for the technology would involve optimising the balance between the number of scans and data quality, potentially through the development of more advanced hardware and algorithms that can extract accurate information from fewer scans without compromising reliability.

6.4.3 Mid-infrared spectroscopy (MIR) instrumentation and maintenance

When selecting the most appropriate instruments for NIRS and MIRS spectroscopy, it is crucial to consider factors such as spectral resolution, which is important when discerning between peaks or features that are very close to each other. This may not be as critical for solid samples but could be for wet or liquid samples. Sensitivity, or the ability of the instrument to detect signals, is also a key factor, along with durability and ease of maintenance, particularly in resource-limited settings where frequent repairs or replacement may not be feasible. High-quality instruments like ATR-FTIR spectrometers offer advanced stabilisation of the IR light source, ensuring consistent performance and longevity, with a typical lifetime of up to 10 years [262].

Additionally, it is important to ensure that the chosen spectrometer can operate under varying environmental conditions, as fluctuations in temperature, humidity and other contextual factors that can affect instrument performance. Instruments equipped with robust software for monitoring and validating performance can help detect and mitigate these effects by providing real-time feedback. In resource-limited settings, the choice of instrumentation will depend on what is locally available. Many countries may already possess different spectrometers in various institutions used for non-entomological purposes (Table 6.4). For instance, spectrometers are widely used in Africa for pharmaceutical and agricultural applications and can be readily repurposed to meet entomological surveillance needs. This approach leverages existing resources, making it a cost-effective solution for implementing advanced vector surveillance programs.

To ensure the quality of spectra data collected by MIRS, the spectrometer must be properly maintained and regularly calibrated. For the ATR-FTIR spectrometers used in the majority of this thesis, the most common maintenance task is replacing desiccant bags. This is important because MIRS is highly affected by environmental micro-climatic conditions such as humidity. Therefore, humidity levels should be regularly checked using the spectrometer software to ensure optimal operating conditions, with a relative humidity (RH) of less than 30% and an absolute humidity of less than $14 g/m^3$ [263]. Maintaining these humidity levels is important because excessive moisture in the air can interfere with infrared spectroscopy by absorbing infrared radiation, leading to increased noise in the spectra. Low humidity reduces this interference, ensuring accurate and reliable spectra data. Additionally, controlling humidity prevents condensation on optical components, which can degrade the performance and longevity of the spectrometer. Moreover, some vendors offer advanced stabilisation of the IR light source in ATR-FTIR spectrometers, with a lifetime of up to 10 years [262]. This demonstrates the durability of the spectrometer, although the IR light source may need to be replaced at the end of its service life.

Table 6.4: Distribution of infrared spectrometers used for entomological and non-entomological purposes across Africa based on a short survey.

Countries	MIRS (n)	MIRS-NIRS (n)	NIRS (n)	Unknown (n)	Total (N)
Benin	-	-	-	2	**2**
Burkina Faso	1	-	1	-	**2**
Kenya	1	-	3	-	**4**
Malawi	-	1	-	-	**1**
Mozambique	1	-	-	-	**1**
Nigeria	1	-	-	-	**1**
Tanzania	2	-	1	-	**3**
Uganda	1	-	-	-	**1**
Total (N)	**7**	**1**	**5**	**2**	**15**

Once maintenance has been performed, the OPUS validation program can be used to carry out validation tests such as Operational Qualification (OQ) and Performance Qualification (PQ) tests [262,263]. The OQ test is typically performed once a year or when a defective optical component, such as the IR light source, has been replaced, to check whether the spectrometer meets the specified performance parameters. These parameters include spectra resolution, sensitivity, energy distribution, and wavenumber accuracy. The PQ test, on the other hand, consists of instrument self-test procedures that test the signal-to-noise ratio and the 100%-line test.

6.4.4 Processing of the infrared spectra data

Upon scanning samples, the spectra data are stored in various formats based on the intended data processing methods. For most NIR and MIRS applications, the default storage format is typically binary file formats, compatible primarily with the default spectrometer software. This format is not directly usable in other open-source software for analysis, but custom-built programs can be developed to convert the computer-readable binary data into a text file format, allowing for broader accessibility and analysis using different software tools [264].

Once collected, infrared data usually require varying degrees of cleaning to eliminate unwanted sections or data points and to ensure appropriate labelling. For instance, noise in MIRS data that originate from atmospheric intrusions like water vapour and carbon dioxide (CO_2) can pose a challenge to obtaining a clean spectrum (Fig. 6.2A & B). Some of this noise can be mitigated by regularly running background measurements to account for any atmospheric interference. However, when atmospheric intrusions are extreme, custom programs can be used to filter the affected spectra out, as well as those with low or no intensity [264]. When atmospheric interferences are extreme, water vapour bands with many narrow peaks appear between 4,000 and 3,400 cm^{-1}, 2,200 and 1,300 cm^{-1}, and 800

cm^{-1} frequencies, while CO_2 strong bands appear at 2,345 cm^{-1}, and weaker bands at 3,650 and 750 cm^{-1} frequencies [59,114]. Moreover, in our experience, spectra with low intensity and all flat spectra with no signals are defined based on the reference band of the plateau between 400 and 500 cm^{-1} frequencies when the average intensity is lower than 0.11 (based on empirical observation), (Fig. 6.2C) [59,114]. Low intensity or no intensity spectra arise when the sample moves slightly on the crystal during scanning, causing a reduction in peak intensity and loss of definition due to a poor signal-to-noise ratio. Therefore, the optimal threshold intensity value needs to be adjusted based on the sample type, as other sample types may require a different threshold. Based on our observations, the threshold should be high enough to maintain good peak definition yet low enough to avoid discarding too many spectra. Since these factors introduce noise to the spectra, it is necessary to drop the affected spectra to ensure quality clean spectra for ML training.

In conclusion, based on the lessons learned in all the different chapters, it can be concluded that infrared spectra data, typically stored in binary formats compatible with default spectrometer software, require custom programs for broader accessibility and analysis using different tools. Additionally, effective data processing involves cleaning to remove low or no-intensity spectra and atmospheric intrusions, ensuring high-quality data for machine learning training.

6.4.5 Using machine learning models for predicting different entomological indicators of malaria based on the infrared spectra data

ML models typically require spectroscopic data pre-processing before training to ensure the data is in the correct format. This improves performance, speeds up training process, and enhances the model's ability to generalise to new, unseen data. Previous studies have used various pre-processing techniques to scale spectroscopy data before training the models, such as mean centring, principal component analysis (PCA), and Savisky-Golay derivative [61,113,114,126,178]. For example, in earlier studies, spectra data were vector-normalised, and the second derivative calculated using Savitzky-Golay algorithm, after which PLS was trained to detect *P. falciparum* in vitro [113,126]. Additionally, other studies pre-processed spectral data by scaling to a mean of 0 and a standard deviation of 1 before training ML models to predict age and species, mosquito host blood-meal source, sporozoite infections, and *P. falciparum* infections in human blood [59,61,114,249,265].

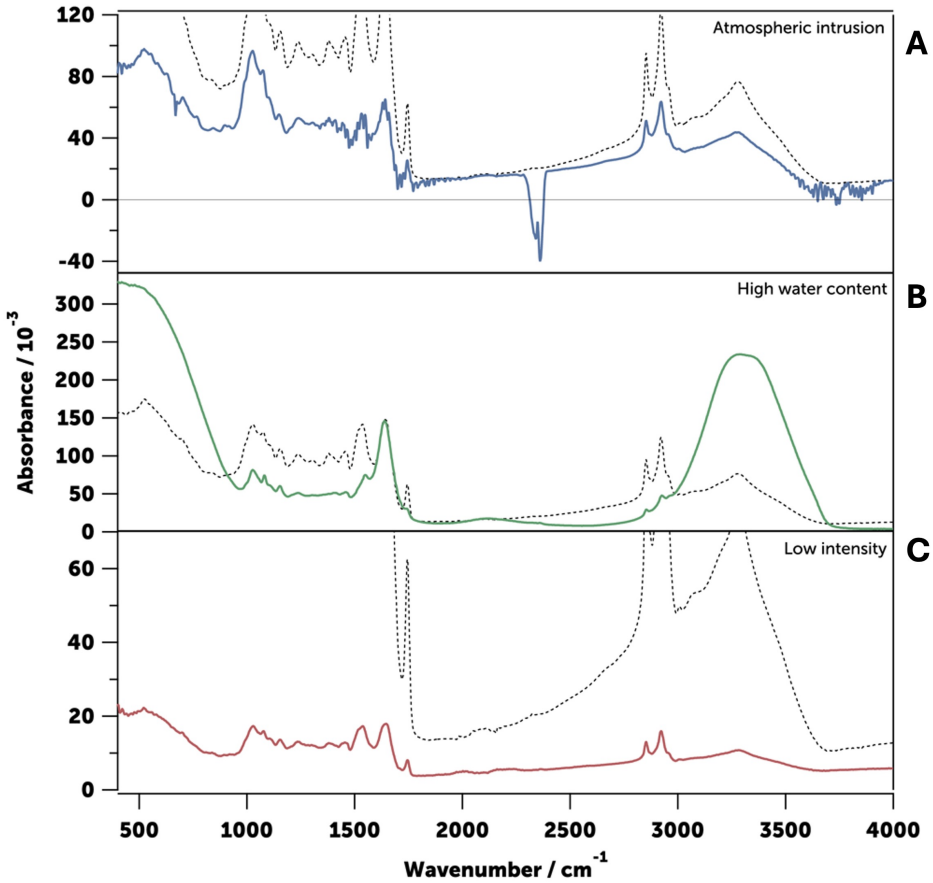

Figure 6.2: Experimental errors leading to noise in mosquito measurement using MIRS. The figure illustrates mosquito spectra with **A**) atmospheric intrusion, **B**) high water content, **C**) poor defined features due to low intensity, compared to the correct measured sample shown with a dashed line [59].

A variety of ML models, including classical/statistical and deep learning models, have been trained and used to predict different entomological and parasitological indicators of malaria transmission from MIR spectra. In Gonzalez Jiminez *et al.*, a linear-based logistic regression model and a tree-based extreme gradient boosting (XGBoost) model were trained to predict mosquito age classes and species, respectively [59]. In this study, the ML models were trained using 17 wavenumbers as features selected from well-defined vibrational absorption peaks and troughs, achieved only 82% classifications accuracy, although they were trained with small number of samples. In Siria *et al.*, a much larger and balanced dataset was used to train CNN model on the entire wavelength of the Mid-infrared spectra, where each wavenumber corresponds to individual vibrational modes [61]. The CNN model, trained on the laboratory data, improved prediction accuracy for mosquito age and species from the same data source [61].

However, inherit variability in the samples limited the generalisability of the models in predicting mosquitoes collected from semi-field or field conditions. To address this, instead of re-training the CNN model entirely on the semi-field or field dataset, transfer learning was used. The weights of the convolutional layers of the CNN trained with

laboratory data were frozen, and only the weights of the dense layers were updated with a subset of new data from either semi-field or field samples [61]. However, training a CNN model is computationally expensive. In a follow-up study, the dimensionality of the spectra data was reduced using PCA and t-distributed stochastic neighbour embedding (t-SNE) to lower computational cost while improving the generalisability of the ML models in predicting mosquito age from two different insectaries [178]. The PCA or t-SNE transformed data can be passed directly to XGBoost or multi-layer perceptron (MLP) models, significantly reducing computational costs [178,244]. However, this alone may not improve generalisability of the ML models, which can only be achieved with transfer learning [61,178]. More evidence of transfer learning is demonstrated with an MLP model trained with laboratory-reared blood-fed *An. arabiensis* spectra data to predict blood-feeding histories of wild-caught *An. funestus* [229]. Table 6.5 highlights the most common models used for analysing MIRS spectra data in malaria surveillance.

Therefore, when training ML models using spectra data, the following factors should be considered: a) data pre-processing; b) in places with high computational capacity, CNN model can be trained; c) in places where computational capacity is limited, dimensionality reduction of the spectra data can be applied, and standard ML models may match the performance of deep learning models [178]; d) transfer learning is important to achieve the generalisability of ML models; e) and for parasite detection in human blood, training a model with the highest parasite concentration can result to robust model capable of predicting malaria infections at different parasitaemia levels, even in the presence of anaemia [265].

Table 6.5: The most common models used to date for analysing MIRS spectra data in malaria surveillance.

Models	Pros	Cons	References
PLS: identifies latent variables (components) by extracting linear combinations of the original predictor to capture maximum covariance with the response variable	• Can handle multicollinearity among predictors • Latent variables can provide insights into the data	• Prone to overfitting, leading to poor generalisation • Not suitable for non-linear relationships	[113, 126]
Logistic regression: predicts the probability of binary or multiple responses based on one or more predictor variables	• Simple, easy to implement, and computationally efficient • Provides insights into important features for prediction • Works well with linearly separable data	• Not suitable for non-linear relationships	[114, 115, 265]
XGBoost: builds an ensemble of decision trees sequentially, where each tree corrects the errors of the previous tree	• High performance • Provides insights into important features for prediction • Suitable for non-linear relationships	• Many hyper-parameters to tune • Computationally expensive for large datasets	[59, 178, 244]
Deep learning (MLP and CNN): uses neural networks with many layers for tasks involving large data and complex patterns	• High performance • Can handle large datasets • Suitable for non-linear relationships	• Requires large labelled datasets for good performance • Computationally expensive • Considered "black boxes" due to difficulty in interpreting weights	[61, 178]

6.5 Key challenges in the application of infrared spectroscopy for malaria vector surveillance

Despite significant advancement, previous research on IR spectroscopy has highlighted several remaining challenges, which must be addressed before this technology can be deployed at scale. These issues include the gaps in interpretability of biological signals, which can be complex and difficult to translate, the lack of generalisability of many machine learning models, and the challenges in field implementation of the infrared spectroscopy and machine learning techniques. Further explanation is provided below and also summarised in Table 6.6.

6.5.1 Limited generalisability of existing algorithms

While ML models trained on spectra data have demonstrated potential for mosquito and parasite surveillance, the inherent variability of samples from different species, diets, environments, and genetic backgrounds may limit the generalisability of these models [204]. For instance, ML models trained using laboratory-generated data may not be transferable to the field, resulting in inaccurate predictions for field-collected data [92, 138, 219]. This challenge can also arise from various sources of variability, including biological differences among samples, technical inconsistencies between users or machines, and model overfitting, where the ML models learn noise rather than meaningful patterns.

Understanding the root causes of these issues is the first step toward resolving them. For instance, biological variability, in predicting mosquito age, could be addressed by targeting more biologically relevant features. For example, training models on biological age rather than chronological age could improve generalisability. Additionally, technical variability introduced by different instruments or operators can be mitigated using standardised protocols and calibration across devices. Overfitting, on the other hand, can be addressed by employing more robust ML techniques, such as regularisation or cross-validation, and by ensuring models are trained on diverse datasets that better represent real-world settings.

Transfer learning has emerged as a promising approach to improve the generalisability of ML models, enabling them to predict mosquito samples regardless of their inherent variability [61, 178, 229]. However, this approach, necessitates repeated sampling from different populations and environments to ensure successful generalisation. Additionally, continuous updates and validation of the models with new data are crucial to maintain their accuracy and reliability in diverse real-world settings. Moreover, a deeper understanding of the field data may also allow for the identification of the key attributes that should be targeted to obtain more generalisable training data. Furthermore, such field collections

should include specific records of metadata, which would be useful for understanding variation in the model results.

A key concern with IR-ML models is the trade-off between accuracy and generalisability. While generalisable models are essential for large scale applications, context-specific models can offer substantial advantages in targeted settings. The choice between these models could depend on the specific goals of the surveillance efforts and the resources available. For large-scale mosquito screening, a slightly less accurate model but more generalisable model could provide consistent and reliable results across various contexts, facilitating broad surveillance efforts. This approach is particularly important in diverse environments, where deploying multiple context-specific models may be impractical.

On the other hand, context-specific models can be invaluable for localised studies or interventions, where high accuracy is essential for precise decision-making. For instance, in regions with unique ecological or epidemiological characteristics, a context-specific model could offer detailed insights that a generalised model might miss.

In making this trade-off, one must consider factors such as the scale of the surveillance program, the variability of the environment being studied, and the criticality of accuracy in decision-making. Ideally, a hybrid approach that combines generalisable models for broad surveillance and context specific models for detailed analysis in key areas could offer best of both worlds.

6.5.2 Gaps in the interpretability of bio-chemical signatures

The ML models rely on the changes in biochemical bonds within lipids, chitin and protein associate with a specific trait, detected through MIR absorption intensities to make predictions or classify biological samples. These models trained on MIRS data leverage these biochemical components to distinguish host blood in mosquito abdomens [114, 229], determine age and species [59, 61, 244, 266], detect the presence of *Plasmodium*-specific proteins in infected human blood, and the biochemical changes in red blood cells following malaria infection [113, 115, 126, 265]. These important biochemical compositions in MIRS are also utilised in other spectroscopic methods, such as matrix-assisted laser desorption ionisation-time of flight mass spectrometry (MALDI-TOF MS) [267–269].

It is evident that ML models are learning from the signals within the X-H stretching region (3,800 - 2,500 cm^{-1}), double bond region (2,000 - 1,500 cm^{-1}), and the fingerprint region (1,500 - 400 cm^{-1}) of the MIR spectra [59, 61, 114, 115, 229, 244, 265]. It is important to note that ML models do not rely solely on individual features to make predictions; rather,

they use combination of features from the MIR spectra wavelength. However, the translation of the features influencing the predictions of ML models is still poorly understood. The recent study observed similarities in signals originating from the fingerprint region of the spectra that influenced ML predictions of three different models [249]. However, differences in signals influencing predictions were particularly noted in the X-H stretching region ranging from 3,500 to 3,000 cm^{-1} frequencies [249]. Therefore, it is still unclear what causes the differences in signals observed in the X-H stretching region for the three models, raising the question whether these signals result from the actual parasite or indicate changes in mosquito cuticular composition due to the presence of the parasite. Moreover, we cannot rule out the possibility that these features or signals picked up by the ML models can be affected by different model iterations.

Furthermore, previous studies have demonstrated that MIRS detect malaria parasite in human blood by picking up signals from parasite byproducts resulting from the digestion of haemoglobin during erythrocytes infection [113, 115, 270]. However, we have observed that methodological variation can significantly impact the performance and generalisability of MIRS-ML for malaria parasite detection. For example, in two separate experiments conducted using the cultured malaria parasite *P. falciparum* NF54 strain, one experiment used normal saline (0.9% NaCl) solution to re-suspend the red blood cells (RBCs), and the resulting ML model failed to predict the same field-collected samples used in the first experiment (Fig. 6.3A) (Mwanga *et al.*, unpublished). In contrast, another experiment used plasma to resuspend RBCs after the removal of the culturing medium, resulting in a laboratory model that successfully transferred to making predictions for the field-collected samples (Fig. 6.3B) [265]. This raises a question of whether MIRS-ML detects actual parasite signals or changes in the blood as a result of parasite invasion of the erythrocytes, a question that requires further investigation.

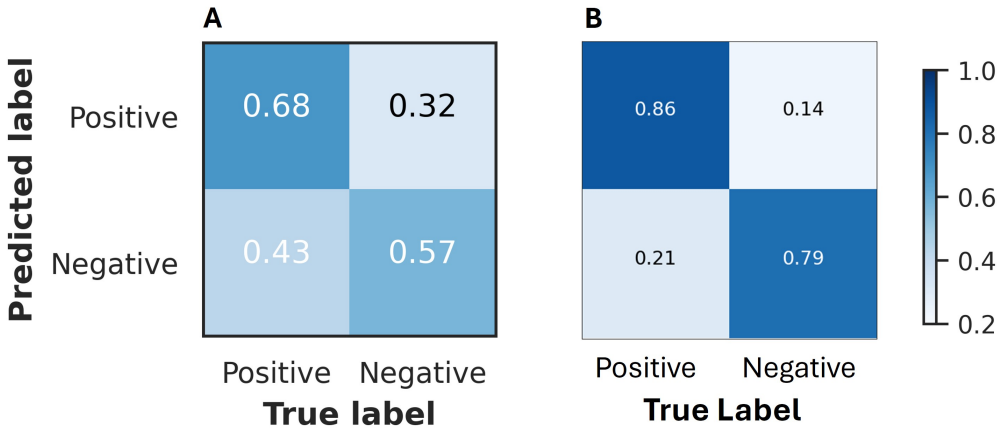

Figure 6.3: Confusion matrix showing the prediction of *P. falciparum* infection in field collected human samples using laboratory-trained ML models. Red blood cells were re-suspended differently after the removal of the culturing medium: **A**) using normal saline (0.9% NaCl), and **B**) using plasma.

Lastly, a critical assessment is necessary to determine the need for supervised versus unsupervised approaches, based on the quality, quantity, and understanding of the available data. In cases where there are large quantities of data, but limited understanding of specific characteristics, deep learning and unsupervised approaches may be more beneficial. These approaches can uncover patterns and structure within the data without prior labels, making them suitable for exploratory analysis and generating hypotheses about the underlying structure of the data. Conversely, when the number of data points is few and there is detailed information about individual samples, supervised models can be more effective. These models can leverage the labelled data to make predictions, and most importantly, to understand the relationship between input features and output predictions.

6.5.3 Implementation of infrared spectroscopy and machine learning

ML has provided a major breakthrough towards creating deployment-ready systems for malaria surveillance. However, further development is necessary, requiring models to be trained on more diverse data to improve the observed accuracies. ML models rely heavily on the quantity and quality of data; more data generally lead to improved performance. Due to data limitations, many MIRS-ML models are still in the development stage. Building deployment-ready ML models that can detect parasite or host-blood meal source in field-collected mosquitoes or parasite in human blood from real field settings requires extensive sampling efforts over an extended period of time. Given that robust ML models are data-hungry, simulating different scenarios that occur in real field settings in laboratory conditions can produce larger datasets. These datasets can be sufficiently large to train deployment-ready ML models. Despite the current developmental stage, with the accuracy achieved in various studies, the question remains whether MIRS-ML should be positioned as a diagnostic or surveillance tool. At present, the technique should serve primarily as a surveillance or screening tool, as the accuracy of 99 to 99.9% for diagnostic has not been achieved.

Obtaining precise and accurate spectral data necessitates the use of MIR spectrometers. However, the lack of proper storage for samples can further impact the quality of the data scanned on the MIR spectrometer. Reliable internet connectivity is crucial for effective implementation of MIRS-ML, as it enables the uploading of datasets and access to cloud-based computational resources. We are currently developing a web app for deploying MIRS-ML models, which underscores the need for stable internet connections.

The effectiveness of surveillance programs can be significantly impacted by the lack of sufficiently trained personnel ML model development. Proper operation of spectrometers requires knowledge, which, if lacking, can lead to inaccurate data collection and analysis. Training ML models demands a deep understanding of both the underlying algorithms and the specific applications to which they are being applied. Without adequately trained personnel, the development and fine-tuning of these models may be sub-optimal, reducing their utility. Regular maintenance of the equipment is crucial to ensure its long-term functionality and reliability. Any neglect in this aspect can result in breakdowns, leading to interruptions in data collection. Effective data management is important to ensure that the information derived from spectra datasets is accurate and consistent. Inadequate storage solutions can lead to data loss or corruption, which can compromise the integrity of the dataset. Therefore, investing in comprehensive training programs for personnel, ensuring they are well-versed in both the technical and practical aspects of training ML models, is essential for success and sustainability of surveillance programs.

The initial cost of procuring advanced spectrometers, coupled with ongoing expenses related to continuous training and data processing can be prohibitive for many organisations in low-resource settings. The substantial upfront investment requires at least £25,000/= for a high-quality spectrometer and associated installation. Additionally, continuous training is crucial for personnel to stay updated with the latest advancement in ML and data analysis techniques. This ongoing education involves not only time but also financial resources, as training programs and professional development workshops can be costly. Furthermore, the costs related to data processing, such as the acquisition of computational resources and data storage solutions, add another layer of expense. These financial challenges can be particularly burdensome for institutions with limited budgets or those in resource-constrained environments.

In terms of applications, commonly used methods for malaria surveillance, such as PCR and ELISA [68,69,75,78,79,213,214,234], can analyse sample by pooling them (i.e., in sizes of 10, 50, 100 etc.) and performing multiplexed reactions. This is particularly important in the areas with low transmission and when resources are limited. The question of whether MIRS-ML can also pool samples requires further investigation. Pooling might be feasible in MIRS-ML, but it will require new spectra data generated from pooled samples, as these would differ from spectra of individual mosquitoes. Additionally, new models would be necessary, as MIRS-ML models trained on individual spectra would likely struggle to generalise to pooled data without modification. These models would need to be trained specifically on data derived from pooled samples to effectively interpret combined signals. However, the spectra signatures of pooled samples would represent a composite of all the mosquitoes in the pool. Disentangling signals from infectious and non-infectious mosquitoes, or mosquitoes of different ages and species, could be challenging and might introduce noise or reduce accuracy. Future investigations are needed to determine whether

signals are gained or lost when samples are pooled in MIRS-ML, given that pooling sacrifice the ability to assess individual mosquitoes. Despite this limitation, MIRS-ML remain a fast and convenient method capable of processing a large number of samples in a short duration. Further research should also explore the best mechanism for pooling samples in MIRS-ML and identify the optimal pool size, as excessively large pools could dilute signals and compromise model accuracy, while smaller pools might not offer significant efficient gains.

Table 6.6: Major challenges to be addressed before the infrared spectroscopy and machine learning techniques can be deployed at scale for malaria vector surveillance.

Challenge	Description	Possible Solutions
Limited generalisability	Variability in samples from different species, diets, environments, and genetic backgrounds limits ML model generalisability. Models trained on lab data may not predict accurately for field data.	Implement transfer learning, perform repeated sampling from diverse populations, and continuously update models with new data.
Gaps in the interpretability of biological signals	Complex biochemical signals in MIR spectra can be difficult to translate, into meaningful predictions. Differences in signals raise questions about their source.	Conduct further research to understand signal sources, improve model interpretation, and explore the biochemical basis of the signals.
Challenges in field implementation of MIRS-ML	Developing deployment-ready ML models requires extensive data from diverse settings. Current models may not achieve diagnostic-level accuracy and may not support sample pooling like in PCR and ELISA.	Generate large datasets in laboratory settings simulating field conditions, focus on surveillance use, and investigate pooling mechanisms for MIRS-ML.

6.6 Conclusion

The application of infrared spectroscopy and machine learning in malaria vector surveillance has shown great promise, offering a rapid, low-cost, and reagent-free alternative to traditional methods. This chapter has summarised the lessons learned during the research activities in the preceding chapters. This reflection has revealed that despite significant advancements, several challenges remain, particularly related to data quantity and quality, generalisability, and scalability, that must be addressed if these technologies are to be used effectively on a large scale. In particular, ensuring the generalisability of ML models is crucial, as the inherent variability in mosquito samples from different species, environments, and genetic backgrounds can limit model accuracy. Transfer learning and continuous model updates with diverse data sets can help improve generalisability, and there are opportunities to make further improvements by enhancing

data quality and coverage. Additionally, the interpretability of biochemical signals in the spectra needs further research to enhance model predictions and understand the sources of variability. For maximum value from the IR spectra, effective data processing is essential for obtaining high-quality spectra data, which involves cleaning to remove low or no-intensity spectra and atmospheric interference. Custom programs may be required to convert binary data into more accessible formats for analysis. Instrument maintenance, such as replacing desiccant bags and performing regular calibration, is also critical to ensure consistent performance. Overall, while MIRS-ML techniques have not yet reached diagnostic-level accuracy, their utility as surveillance tools is clear. Future research should focus on creating deployment-ready systems, including investigating pooling mechanisms for sample analysis and generating diverse datasets to train robust ML models. Addressing these challenges will be key to maximising the potential of infrared spectroscopy and machine learning for effective malaria vector surveillance.

Chapter 7

General Discussion

7.1 Overview of the main findings

The overall aim of this thesis was to demonstrate that mid infrared spectroscopy coupled with machine learning (MIRS-ML) can provide high-throughput and accurate assessments of mosquito age, blood-feeding histories, and detection of *Plasmodium falciparum* in field-collected mosquitoes. Additionally, the thesis discusses key lessons learned and potential future directions for the use of MIRS-ML in malaria surveillance. Insights gained through this work mark an important step towards creating a deployment-ready system for malaria vector and parasite surveillance, particularly in malaria-endemic and low-resource settings where routine surveillance techniques are costly, labour intensive and time consuming.

Prior to this work, near-infrared (NIR) spectroscopy had already proven useful for mosquito age-grading and species identification [60, 94, 147, 176]. Furthermore, early evaluations of mid-infrared (MIR) spectroscopy had suggested that it could offer significantly greater resolution, enabling the detection of finer biological signals [59, 61]. At that time, robust desktop equipment for these techniques was available, suggesting that these tools could be deployed beyond controlled laboratory environments. However, critical questions remained: Would MIRS-ML be used solely as an age-grading and species identification tool? Could it be effectively deployed in the field? What additional entomological indicators could it track?

Before starting the work for this PhD, I had already taken the first steps toward addressing these questions by demonstrating that MIRS-ML was versatile enough to measure additional entomological indicators [114]. Notably, I had shown, using laboratory-generated samples, that MIRS-ML could accurately assess human blood indices, offering a valuable method for studying mosquito blood-feeding histories [114]. This, in turn, would enable the determination of anthropophily, which is the degree to which a vector bites humans over other vertebrates. This is considered a direct measure of how suitable a mosquito species could be for transmitting human diseases and therefore a critical factor in evaluating transmission abilities of vectors [41, 136].

Building on this foundation, a further objective of this thesis was therefore to explore MIRS-ML as a multipurpose tool, evaluating its suitability for tracking multiple entomological traits in field mosquito populations. The research presented here therefore represents the most comprehensive assessment to date of the applications of infrared

spectroscopy and machine learning for tracking essential entomological indicators of malaria. Taken together, these findings represent a significant advancement towards the creation of a one-stop, deployment-ready platform for malaria vector and parasite surveillance, particularly in malaria-endemic and resource-limited settings.

This general discussion synthesises the key findings from each chapter, highlights the potential of MIRS-ML in malaria surveillance, reflects on its limitations, and identifies key questions for future research.

7.2 Transfer learning and dimensionality reduction to mid-infrared spectra data to improve the transferability and generalisability of mid-infrared spectroscopy and machine learning (MIRS-ML) based predictions for mosquito ages

In the early stages of our research, one of the significant challenges was the need to develop context-specific models to answer particular questions or estimate specific attributes within different settings and circumstances. These models, while effective in their respective laboratory environments, often lacked generalisability, limiting their utility outside the specific settings where they were developed. One solution was to collect new data from each new setting and train entirely new models. However, this approach was labour-intensive and time-consuming. A more efficient option emerged through the use of pre-existing models that could be re-calibrated using small subsets of data from the new target environments – a technique known as transfer learning.

Recognising the potential of transfer learning, this thesis began with exploring whether transfer learning approaches could improve the generalisability of both deep learning and standard machine learning models in predicting mosquito age classes across different rearing conditions, such as those in various insectaries. By fine-tuning pre-trained models with limited, context-specific data, this method holds the promise of significantly reducing the need for extensive retraining while increasing the applicability of models to diverse environmental settings. This exploration represents a critical step toward creating more adaptable, scalable tools for malaria vector surveillance.

Machine learning models were trained with data from Ifakara and evaluated with 1,635 spectra from Glasgow-reared mosquitoes. The findings demonstrated that transfer learning can improve the generalisability of the ML models in predicting mosquito age classes across different locations; by correcting differences in data distribution between

training and evaluation datasets. Furthermore, reducing the dimensionality of the spectral data reduced computational costs (i.e., computing time) when training the ML models.

This study showed that performing transfer learning using only 2% of the spectra from the target domain (Glasgow-reared mosquitoes, 33 of 1,635) as well as dimensionality reduction resulted in the improved generalisability of both deep learning and standard ML models, achieving an overall accuracy of 98%. The expectation was that all models with transfer learning applied would outperform the baseline models without transfer leaning in generalising predictions, as previously demonstrated [61,148]. Furthermore, dimensionality reduction like PCA and t-SNE were used on different occasions to reduce noise and redundant features in the spectra, as well as to decrease computational time when training the ML models. However, the findings indicate that dimensionality reduction alone cannot achieve the generalisability of the ML models but can when used in conjunction with transfer learning. The use of PCA improved model stability by projecting data into a lower-dimensional scale while preserving the original distance between the data points [124]. On the other hand, t-SNE failed to improve generalisability of the ML models due to its probabilistic nature and a non-convex cost function [140], resulting in different outputs from multiple runs, which may not preserve the original distance between the data points.

The improvement observed in model performance on the target population following the addition of 2% of data is attributed to the transfer learning process, which involves fine-tuning a pre-trained model on new data. This process leverages the knowledge encoded in the pre-trained model, enabling faster convergence and better generalisation with limited new data. To isolate the contribution of transfer learning from mere data addition, future studies could compare performance by retraining a model from scratch with 2% added data versus fine-tuning a pre-trained model on the same data. While computational efficiency is an important advantage of transfer learning, the process itself is crucial as it allows the model to adapt to the target population by building on previously learned features rather than starting from zero. This is particularly important in cases with small datasets, where retraining a model from scratch often fails to achieve comparable performance due to the lack of sufficient training data to develop robust feature representations.

Therefore, transfer learning approaches became standard in all subsequent work I did, offering a practical solution for model training in new settings that takes into account the high sensitivity of mosquito development to its environment, and would otherwise require extensive data collection. This approach allows users to build on existing models rather than starting training from scratch each time.

7.3 Classification of the epidemiologically relevant age of malaria vectors

An. funestus is not only the main malaria vector in Tanzania but also serves as an excellent model for studying *Anopheles* vectors in general, due to its relatively higher survival rates compared to other vectors, such as *An. arabiensis* [135,271], making age-grading this species particularly relevant. Several factors make age-grading in *An. funestus* unique. In addition to being one of the longest-living malaria vectors in Tanzania, it also exhibits significantly delayed maturity rates. It takes longer to mature sexually (Hape *et al.*, Unpublished); has nearly twice the duration in its aquatic stages compared to contemporary vectors [272]; requires almost twice as much time to mate (Hape *et al.*, Unpublished); and consistently features higher parity rates in the field. Moreover, it shows stronger resistance to insecticides than other vectors in Tanzania, allowing it to survive multiple exposures to insecticidal interventions [135]. Therefore, after age-grading had been demonstrated in *An. arabiensis*, *An. gambiae*, and *An. coluzzii* [61], it was clear that the most appropriate model for age-grading in *Anopheles* should include *An. funestus*.

Chapter 3 presented a study that used 2,084 spectra data points to train ML models to classify the epidemiologically relevant age groups of *An. funestus* mosquitoes. These mosquitoes were reared from wild larvae using water from their natural habitats but under controlled laboratory conditions. One of the key concerns with previous applications of MIRS-ML based approaches for entomological assessment is the lack of validation for wild-caught malaria vectors in field settings, with few exceptions in semi-field and field setting [61]. In this study, *An. funestus* larvae were collected from various villages and breeding habitats to account for genetic variation, differences in larval food sources, and microbiome diversity, maintaining some characteristics of natural ecosystems.

The ML models successfully distinguished between young *An. funestus* females (1-9 days old) and the older ones (10-16 days old) based on the MIR spectra, which is likely to reflect the varying biochemical composition of the mosquito cuticles. The study did not attempt to classify mosquitoes based on exact chronological ages due to insufficient sample size. Instead, the selected age classes represent a typical epidemiological distinction relevant to malaria parasite transmission, which requires a vector to be at least 10 days old under optimal conditions [20]. The success of this analysis and the high accuracies obtained indicate the potential of this approach for predicting key mosquito attributes in field settings. Although this was the first demonstration of the effectiveness of this technique for predicting the age of *An. funestus* mosquitoes, the MIRS-ML approach has been widely demonstrated for predicting indicators such as age, blood meals in other *Anopheles* species [59,61,114,178].

Building on the findings from the previous chapter, dimensionality reduction was used to reduce noise and redundant features in MIR spectra. Initially, this chapter relied on the capabilities of XGBoost model to select the most dominant features (e.g. by selecting the top 100 features) which the model had used for prediction. These important features were mostly associated with proteins, with the most influential peak (1,700 cm-1) corresponding to the amide bond of proteins. This suggests that the model learns from protein-based biological traits that vary with the mosquito age. Additionally, PCA was applied to reduce the dimensionality of the spectra data by projecting them into eight principal components [178], resulting in prediction accuracy comparable to the model trained with top 100 features extracted from the XGBoost model. This indicates that ML models are more accurate when trained with fewer features that explain more variations in the data, rather than many redundant features that introduce noise. Furthermore, reducing the dimensionality of the spectra data can reduce the computational requirements for training ML models, which is especially beneficial in areas with limited computational capacity.

7.4 Detection of blood meal sources in field-collected mosquitoes

Malaria transmission between vertebrate hosts is facilitated by the blood-feeding behaviour of mosquitoes, which allows pathogens to establish and be transmitted by the mosquitoes. The HBI is critical for understanding vector transmission dynamics, especially in high-risk malaria areas. HBI is one of the strongest indicators of malaria vector capacity and is considered an important metric in assessing malaria vector transmission across various settings [41]. Afro-tropical malaria vectors are especially dangerous due to their high anthropophily [27]. These vectors pose significant risks in regions with frequent human-mosquito contact.

In my previous work, I used laboratory-reared *An. arabiensis* to identify host types with known identities [114]. Building on that study, I moved to field validation, where the host range, genetics, and digestion times were unknown, to assess real-world blood-feeding histories. This transition was crucial for providing a more realistic assessment of host preferences, which is essential for understanding transmission dynamics. For this study, mosquitoes scanned with MIRS to assess blood-feeding histories excluded gravid individuals. Unlike the other chapters on age and infection status, where mosquitoes were collected using host-seeking traps [227], this study relied primarily on resting mosquitoes, as they tend to rest after feeding. To ensure a comprehensive analysis, I expanded collections to multiple locations, including animal shelters, outdoor areas, and indoor environments.

Given the known feeding behaviours in Kilombero [128, 190, 273], I anticipated finding significant variations in host preferences across different ecological settings.

Chapter 4 of this thesis demonstrated the first-ever field application of the MIRS-ML approach for predicting the blood-feeding histories of malaria vectors in rural Africa. This chapter further demonstrated the transferability of the laboratory-trained MIRS-ML models to classify host blood meals in field-collected samples using transfer learning techniques. PCR served as the ground truth to determine the actual blood-feeding histories of the field-collected mosquitoes, with a total of 1,854 blood-fed *Anopheles* mosquitoes being examined.

Among these field-collected mosquitoes, the majority of mosquitoes were confirmed to have fed on either human or bovines. Consequently, binary classifiers were trained for host blood-meal prediction. This chapter demonstrated the capability of MIRS-ML models to classify mosquito blood-meal sources with high accuracy using a well-balanced set of 338 spectra data from field samples (169 human-fed and 169 from bovine-fed mosquitoes). This demonstrates a significant opportunity to deploy MIRS-ML for estimating HBI, thereby extending the capability of infrared-AI-based systems already proven effective for assessing other entomological attributes such as age and species [204].

Comparatively, earlier work focused on age-synchronised, laboratory-reared, blood-fed *An. arabiensis* achieved a classification accuracy of ~98% in predicting four mosquito host blood-meal sources (i.e., bovine, human, goat, and chicken) [114]. That study was limited to mosquitoes that were only 6-8 hours post feeding. In contrast, the field mosquitoes used in this chapter represented a broader range of age groups and natural variation in blood-meal digestion stages, suggesting the potential of MIRS-ML for realistic field surveillance, even when the actual time of blood-feeding and digestion stages is unknown upon sample collection.

A major achievement highlighted in this chapter is the successful demonstration of the transferability of laboratory-trained models to field samples through the application of transfer learning. Initially trained using spectra data from blood-fed *An. arabiensis* [114], the base laboratory model was updated by incorporating a small subset (n = 100, with 50 samples each from humans and bovine blood-fed *An. funestus* spectra) of field-collected data. With the application of transfer learning, the updated model successfully predicted the blood-meal sources of field-collected *An. funestus* with a classification accuracy of 90%. This finding indicates that MIRS-ML can be extended to detect blood-meal sources of different Afro-tropical malaria vectors, suggesting that species would not be a confounding factor. Furthermore, transfer learning enables prediction to bridge the gap between laboratory and field samples from different locations, ensuring that the origin of the sample does not influence the accuracy of the results. Moreover, the blood content in laboratory-reared mosquito is comparable to that of field-collected mosquitoes; for

instance, the digestion process in mosquito that feed on human host in the laboratory is likely to be the same as that of a field mosquito feeding on a human. Therefore, blood meal models are likely to be more easily transferable between laboratory and field settings compared to other models, such as those predicting age or infection status, which ted to differ more significantly between insectary and field conditions. Additionally, the age of field-collected mosquitoes is unlikely to be a confounder, as similar transfer learning approaches can mitigate this issue. This chapter further contributes to the growing body of evidence that utilising transfer learning can significantly enhance the generalisability of ML prediction for entomological attributes of malaria transmission, as demonstrated for age and species across different countries and laboratories [61, 178].

Finally, this chapter demonstrated that the transferability of laboratory-trained models to field conditions not only enhanced the classification accuracy for blood-fed mosquitoes collected from the field, but also improved the precision in estimating the HBI compared to the ground truth PCR method. This indicates that the technique has the potential to be a reliable method for estimating HBI, capable of generalising HBI estimations in field-collected mosquitoes as effective as PCR. Therefore, it can provide valuable information to national malaria control programs regarding the feeding preferences of blood-fed mosquitoes.

7.5 Detection of *Plasmodium*-infections in field-collected mosquitoes

In chapter 5, this thesis demonstrated the field application of MIRS-ML for detecting *Plasmodium falciparum*-infectious *Anopheles* mosquitoes. This was achieved by collecting and analysing the MIR spectral signatures from the heads and thoraces of wild-caught *An. funestus* females in rural Tanzanian villages. These findings were validated using ELISA or PCR to confirm the presence of *P. falciparum* sporozoite, establishing a reliable 'ground truth' for model training. The findings showed that MIR spectral analysis can distinguish between infectious and non-infectious mosquitoes with accuracies exceeding 90% in some instances. These findings are consistent with studies that used NIRS frequencies in laboratory settings, which reported the detection of *P. falciparum* sporozoite infection in *An. gambiae* mosquitoes [93], and *P. berghei* sporozoite infection in *An. stephensi* [215]. Earlier models trained on NIRS failed to identify mosquitoes infected with wild-strain parasites from asymptomatic malaria carriers, possibly due to limitations in the training dataset [219]. Models trained on MIRS, which provide clearer peaks with richer biochemical information appear to perform better [98, 174]. This enhancement suggest that MIRS-ML can also detect infections in wild-caught mosquitoes, a capability not fully realised with NIRS models in previous studies.

Furthermore, the study found that the model trained with ELISA infection dataset had limitations in predicting samples screened by PCR. Instead, the ML model trained with the PCR infection dataset demonstrated a robust ability to generalise its prediction in classifying infectious and non-infectious wild-caught *An. funestus* mosquitoes screened by ELISA. This generalisability can be attributed to the sensitivity of PCR in detecting even low sporozoite numbers in mosquitoes [224], thereby training set might have been 'more' reliable, thus enhancing the performance of MIRS-ML models.

The biological requirement that mosquitoes must exceed a certain age threshold (i.e., over 10 days) to become infectious, due to the requisite extrinsic incubation period for the parasite [20], introduce potential age-related biases in detection efficacy. However, in this study, mosquito age did not significantly interfere with the prediction of infectious and non-infectious *An. funestus*. Analysis demonstrated that the machine learning models adeptly identified signals indicative of infection regardless of age, thus negating age as a significant confounding variable in our study.

Lastly, this study contributes to the expanding body of knowledge showcasing the potential of MIRS-ML based approaches for malaria vector surveillance. The use of MIRS-ML in assessing key entomological parameters such as age, species identification, and blood-feeding patterns of mosquitoes has been well documented [59,61,114]. The outcomes of our study suggest that MIRS-ML could serve as a versatile platform, enabling the interpretation of infrared scans to ascertain not only the species and age of mosquitoes, factors critical to their potential as malaria vectors, but also their blood-feeding history on humans or other vertebrates, and their infection status with malaria parasites. Although this advancement was tested exclusively on *P. falciparum* and *An. funestus*, this advancement represents a significant step forward in developing MIRS-ML for malaria surveillance.

7.6 Key lessons learned from infrared-based entomological and parasitological studies so far, and the potential future directions

Building on the findings from chapters two, three, four and five, chapter 6 concludes this thesis with a literature review that explores the lessons learned from mid-infrared based entomological and parasitological studies and the potential future directions of research using MIRS-ML technique. The literature reveal that applying MIRS-ML in malaria vector surveillance shows great potential, offering a cost-effective and regent free alternative to commonly used methods, especially in resource-limited settings.

Effective data processing remains essential for obtaining high quality spectra. This includes cleaning the data to remove noise caused by low or no-intensity spectra or atmospheric interference, such as water and carbon dioxide. Custom programs may be required to convert binary data formats into more accessible text formats for broader analysis. Instrument maintenance through regular calibration and reducing humidity levels is also essential for consistent performance in MIRS.

While significant progress has been made in predicting mosquito age, species, blood meal sources, sporozoite infection and parasite infection in human blood, challenges in generalising ML models persist. The MIRS-ML models are indeed learning true signals rather than just noise. This is supported by several factors: First, the models demonstrate consistent performance across experiment in controlled laboratory environments, where they have shown high accuracy in predicting the aforementioned attributes. The consistence of these results across different settings and mosquito species suggests that the models are learning meaningful patterns from the spectra. Second, studies have validated MIRS-ML model predictions against established method such as PCR and ELISA, which serve as 'ground truth' for model training [115, 229, 249]. The high agreement between MIRS-ML predictions and these widely used methods further support the idea that the models are detecting real biological signals. Moreover, MIRS-ML models have been able to highlight specific spectral features linked to biological process such as protein degradation or parasite presence [114, 115, 229].

Despite learning biological signals, transfer learning is often necessary due to inherit variability in real-world data. Mosquitoes from different locations may have diverse genetic backgrounds, environmental exposures, physical conditions, diets, and age distributions, all of which can affect the spectral signatures. Given the complex and variable nature of mosquito populations and environmental factors, transfer learning become essential to maintain model accuracy across different datasets. Additionally, improving the interpretability of biochemical signals within the spectra could further enhance model predictions and clarify the sources of variability in spectra data.

For wider deployment-ready system, more data diverse data are needed, which is currently a limitation. Field collections is resource-intensive, and laboratory-generated datasets simulating field conditions may offer a practical solution to fill this gap.

7.7 Limitations of the study and next steps

This thesis has demonstrated significant progress toward developing a deployment-ready MIRS-ML based-approach for high-throughput and accurate assessments of mosquito age, blood-feeding history, and *Plasmodium falciparum* detection in field-collected mosquitoes.

However, several challenges and limitations require further consideration and investigation. These limitations, which are linked to specific chapters, are discussed in the individual studies but are summarised here to provide a comprehensive overview.

In chapter 2, the study focused on a single species, *An. arabiensis*, originating from two insectaries, to evaluate the impact of transfer learning and dimensionality reduction on the spectra data in improving the generalisability of ML models. While the results showed improved model generalisability, relying on one specie may not adequately represent other Afro-tropical malaria vectors. Future research should test the technique with more diverse samples from different laboratories, field settings, and mosquito species to enhance the predictive capacity of the model. Additionally, this study tested 2% and 5% transfer learning ratios, while Siria *et al.*, used ~10% [61]. Although a 2% transfer learning ratio was sufficient for achieving generalisability, the optimal ratio for MIRS-ML in mosquito and parasite surveillance remain undetermined and warrants further investigation.

In chapter 3, the adult mosquitoes were fed only a 10% sugar solution and collected based on chronological age. Although MIRS-ML models successfully classified epidemiologically relevant age categories, future studies should incorporate other physiological factors such as blood-feeding and oviposition, which could help build more robust age classification models [60, 90]. This could be achieved by maintaining mosquitoes in a semi-field or insectary setting, where upon emergence from pupae, a subset of mosquitoes is collected daily. The remaining mosquitoes would be allowed to undergo natural physiological processes such as blood-feeding, gonotrophic cycles, and oviposition until the population is naturally depleted. This process would be repeated across multiple cohorts until the desired sample size is achieved, providing a robust dataset for training and validating models that reflect real-world biological conditions. Additionally, future research should explore the different ageing rates, the correlation between ML-classified age categories and malaria transmission epidemiology, including infection rates in mosquito and prevalence in human populations. Furthermore, although the mosquitoes were collected in the field as larvae and reared in an insectary using water from their natural habitats to mimic field conditions, this approach does not account for microbial changes that may occur during the transition from field to insectary environments.

In chapter 4, in field settings – mosquitoes feeding on multiple host – pose a challenge in field applications. While PCR effectively detect mixed blood meals, it is unclear whether MIRS-ML can achieve the same, making this an area for further research. The ML model in this study were trained only on mosquito that had fed on human and bovines, limiting their application in real-field settings where other potential hosts, such as chicken, pigs, goats, and dogs, are present. Future ML models should include a broader range of hosts to create a fully deployed system for detecting mosquito blood meal sources. Although

extensive field sampling may be required to obtain well-balanced datasets which also not all of them will amplify with PCR or ELISA, laboratory-generated data simulating various hosts near human dwellings could be used, with transfer learning applied for field predictions. Moreover, This study did not assess how the model (i.e. trained with human and bovine blood-fed samples) performs when predicting non-human or non-bovine blood-fed samples. It would be valuable to investigate the model's predictions for such samples and determine whether their predictions probabilities are systematically lower than those for actual human or bovine samples. This could include using prediction probabilities to establish a threshold that restricts the model from classifying samples as human or bovine blood-fed if they are likely to belong to other hosts.

In chapter 5, the study focused on screening individual mosquitoes for sporozoite infection. In low transmission settings, however, sporozoite infections are rare [39]. PCR and ELISA can pool mosquito samples to reduce operational costs and time, but it is yet to be determined if MIRS-ML can perform pooled testing. Furthermore, the study only focused on *P. falciparum* and did not assess co-infections with other *Plasmodium* species. Since parasite-specific proteins, such as circumsporozoite (CS) protein and the thrombospondin-related adhesive protein (TRAP) [220,221], or immune responses elicited by mosquitoes as a result of parasite infection, may influence the biochemical characteristics of infectious and non-infectious mosquitoes, it is possible that MIRS-ML could detect any *Plasmodium* species. This suggests that the presence of multiple *Plasmodium* species co-circulating in a region might not pose a significant challenge for the real-world application of MIRS-ML. However, it remains unknown whether differences among species could influence model performance. Future studies are needed to confirm whether the model is equally sensitive to infections by different *Plasmodium* species and to assess its ability to handle co-infections. Additionally, to understand the model's sensitivity to *P. falciparum*, further research could involve training models using data from infections exclusively caused by *P. falciparum* and other specific *Plasmodium* species, allowing a comparison of predictive performance. Alternatively, deeper exploration of the biological basis of the models – such as identifying which biochemical features in mosquito spectra correspond specifically to *P. falciparum* infections – could provide insights into improving model sensitivity and specificity.

There are common limitations affecting all chapters (CHAPTER 2, 3, 4 & 5). The availability of sufficient data for training ML models remains a significant constraint. Robust ML models rely on large, high-quality datasets, and future efforts should focus on generating additional and more diverse laboratory data to develop more robust and deployment-ready systems. Furthermore, the generalisability of the models is a key challenge, as they need to be applicable across diverse settings and mosquito species to be truly effective. Additionally, understanding the biological signals captured by ML models remains difficult, which hinders the interpretability of the data. Lastly, practical issues related to field implementation need to be addressed to ensure that these models can be

effectively deployed in real-world scenarios – these challenges are discussed in detail in chapter 6.

Another key limitation is that the different variables – mosquito age, blood-feeding histories, and infection status – were evaluated separately in different experiments and with different mosquitoes. While this was necessary to validate the technique's suitability for each indicator, the next crucial step will be to test all these indicators on the same mosquito. This integration could significantly enhance the workflow's efficiency and utility. One potential solution could involve the use of robotics or automated bench-top systems to streamline the process and reduce menial labour. This may allow for faster and more efficient data collection, making the entire workflow more practical for large-scale deployment. Future studies should investigate the feasibility of integrated systems capable of assessing multiple indicators simultaneously.

7.8 Conclusion

The work presented in this thesis has been a significant step in validating the potential of mid-infrared spectroscopy coupled with machine learning (MIRS-ML) for high-throughput, accurate assessments of mosquito age, blood-feeding histories, and *P. falciparum* infections in field-collected mosquitoes. In totality, it represents a significant advancement from prior work with both near-infrared and mid-infrared spectroscopy approaches – ultimately showing that MIR can offer greater resolution and versatility. The incorporation of transfer learning was particularly crucial in enhancing the generalisability of machine learning models, allowing predictions to be applied across different mosquito populations and environments. These advancements mark important progress toward the development of a deployment-ready system for malaria vector and parasite surveillance, offering a scalable and cost-effective solution for resource-limited settings. However, further research is needed to address challenges in data availability, model generalisability, and field implementation to fully realise the potential of MIRS-ML in large-scale malaria surveillance programs. While each entomological indicator was tested separately in this study, without assessing multiple variables in the same mosquito, the results consistently suggest the potential of MIRS-ML as a multipurpose tool for tracking a wide range of mosquito attributes. Future research should therefore focus on integrating these capabilities into a unified system to be more readily applicable for understanding disease transmission dynamics and evaluating vector control interventions in resource-poor settings. Realising this vision will mark a transformative advancement in field-based entomological surveillance systems, with the potential to revolutionise malaria control and elimination efforts in Africa.

Key Messages

Technical summary

Effective malaria surveillance and control require a thorough understanding of key biological traits, including preferred blood-hosts, infection rates, survivorship, and age distribution. Current methods such as polymerase chain reactions (PCR), enzyme-linked immunosorbent assays (ELISA), and dissection are labour-intensive, time-consuming, and require expensive reagents, making it difficult for routine surveillance, especially in low-resource settings. Advances in infrared spectroscopy and machine learning (MIRS-ML) may offer a faster, cost-effective alternative for predicting mosquito age and species, identifying blood source and detecting pathogens like *Plasmodium*. My PhD aimed to validate and extend MIRS-ML for malaria vector surveillance by improving its generalisability, particularly for predicting mosquito age, blood-feeding histories, and *P. falciparum* infection status.

By applying transfer learning and dimensionality reduction, I was able to significantly improve MIRS-ML model accuracy in classifying key entomological and parasitological indicators of malaria transmission. This research represents an important step toward creating a scalable, field-deployable system to enhance malaria surveillance and intervention monitoring.

Lay summary

Malaria surveillance is essential for understanding disease transmission, planning interventions, and assessing their effectiveness in endemic regions. Traditional surveillance methods are time-consuming and expensive, making them difficult for large scale implementation. My PhD focused on demonstrating that MIRS-ML can serve as a high-throughput, cost-effective alternative for assessing key malaria transmission indicators, such as mosquito age, blood-feeding behaviour, and infection status. Specifically, I applied MIRS-ML to predict mosquito age in various species, detect blood-meal sources and assess *P. falciparum* infection in *An. funestus*. This research provides a significant step toward scaling MIRS-ML for broader in malaria control programs.

Key findings and messages

1. **MIRS-ML for malaria surveillance:** MIRS-ML proved highly effective in predicting mosquito age, blood-feeding histories, *P. falciparum* infection in field-collected mosquitoes. The method has great potential for broadening its use to all Afro malaria vectors, making it a valuable tool for surveillance.

2. **Data pre-processing is important:** Effective data cleaning and instrument maintenance are essential for obtaining high-quality spectra. Removing low-intensity signals and atmospheric noise improves the model's performance and reliability. Regular instrument calibration and customised data processing are vital for achieving consistent results.

3. **Transfer learning improves the generalisability:** Transfer learning significantly improves model accuracy across different locations by adapting laboratory-trained models for use with field data. This technique allows for better predictions of mosquito age and blood-feeding history, even when environmental and biological condition vary. Optimal transfer learning ratios also need to be established for mosquito and parasite surveillance.

4. **Dimensionality reduction:** While dimensionality reduction alone does not guarantee model generalisability, when combined with transfer learning, it can reduce computational requirements and improve efficiency . Probabilistic methods like t-SNE did not improve the stability compared to PCA, which is more reliable for consistent results.

5. **Fewer, more relevant features enhance model performance:** Model trained with fewer, high-variance features outperform those with many redundant ones. This reduction also lowers computational time, making it more practical for real-field applications.

6. **PCR is preferable as the "ground truth" for training models for sporozoite detection:** Models trained on PCR data proved more accurate than those trained on ELISA data for detecting *P. falciparum* infection in mosquitoes. This suggests that PCR, with its sensitivity, should be the "ground truth" for training future ML models.

7. **Species and environmental diversity needed:** Although transfer learning improved generalisability, future studies should include more diverse mosquito species and environments to further enhance the predictive capacity of MIRS-ML.

8. **Physiological factors for age prediction:** Future research should incorporate additional physiological factors, such as blood-feeding and oviposition, to improve

the robustness of mosquito age classification. Moreover, understanding the relationship between mosquito age and malaria transmission is critical for effective vector control.

9. **Data Limitations:** A key challenge in this study was the limited data available for training ML models. More extensive and diverse datasets will improve model robustness and generalisability. Additionally, challenges such as mixed blood meals, pooled testing, and co-infections need to be addressed in future studies to make MIRS-ML more applicable in the field.

Personal learning

This PhD program has equipped me with a versatile skill set, enhancing my scientific and academic writing, project management, data analysis, and molecular biology expertise. I have also developed a deeper understanding of machine learning and statistics, which have been essential for analysing complex datasets. Perhaps most importantly, I have learned the value of collaboration and knowledge exchange, which has been critical for advancing research and making a meaningful impact in malaria control.

Bibliography

[1] T. M. C. and S. L. R., "Climate Change and Vectorborne Diseases," *New England Journal of Medicine*, vol. 387, pp. 1969–1978, 11 2022.

[2] WHO, "World Malaria report 2023," tech. rep., WHO, Geneva, 2023.

[3] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[4] WHO, "Disease Outbreak News; Dengue – Global Situation," 2024.

[5] E. A. Cromwell, C. A. Schmidt, K. T. Kwong, D. M. Pigott, D. Mupfasoni, G. Biswas, S. Shirude, and et al., "The global distribution of lymphatic filariasis, 2000-18: a geospatial analysis," *The Lancet Global Health*, vol. 8, pp. e1186–e1194, 9 2020.

[6] WHO, "World malaria report 2022," tech. rep., WHO, Geneva, 2022.

[7] Ministry of Health Tanzania Mainland, M. o. H. Zanzibar, N. B. o. Statistics, and et al., "Demographic and Health Survey and Malaria Indicator Survey 2022 Key Indicators Report," tech. rep., Ministry of Health, 2023.

[8] Tanzania Commission for AIDs, Zanzibar AIDS Commission, National Bureau of Statistics, and O. o. t. C. G. Statistician, "Tanzania HIV/AIDS and Malaria Indicator Survey, 2007-2008," tech. rep., Ministry of Health, Dar es Salaam Tanzania, 2007.

[9] F. M. Mashauri, S. M. Kinung'Hi, G. M. Kaatano, S. M. Magesa, C. Kishamawe, J. R. Mwanga, S. E. Nnko, R. C. Malima, C. N. Mero, and L. E. Mboera, "Impact of indoor residual spraying of lambda-cyhalothrin on malaria prevalence and anemia in an epidemic-prone District of Muleba, North-western Tanzania," *American Journal of Tropical Medicine and Hygiene*, 2013.

[10] N. Protopopoff, J. F. Mosha, E. Lukole, J. D. Charlwood, A. Wright, C. D. Mwalimu, A. Manjurano, F. W. Mosha, W. Kisinza, I. Kleinschmidt, and M. Rowland, "Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two fact," *The Lancet*, vol. 391, pp. 1577–1588, 4 2018.

[11] National Malaria Control Programme (Tanzania), "National Malaria strategic Plan 2021-2025: Transitioning to Malaria Elimination in Phases," tech. rep., National Malaria Control Programme, 2020.

[12] 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*, 2015.

[13] J. Nájera, M. González-Silva, and P. L. Alonso, "Some lessons for the future from the Global Malaria Eradication Programme (1955–1969)," *PLoS Med*, vol. 8, p. e1000412, 2011.

[14] M. F. Finda, I. R. Moshi, A. Monroe, A. J. Limwagu, A. P. Nyoni, J. K. Swai, H. S. Ngowo, E. G. Minja, L. P. Toe, E. W. Kaindoa, M. Coetzee, L. Manderson, and F. O. Okumu, "Linking human behaviours and malaria vector biting risk in south-eastern Tanzania," *PLOS ONE*, vol. 14, p. e0217414, 6 2019.

[15] A. Monroe, S. Moore, H. Koenker, M. Lynch, and E. Ricotta, "Measuring and characterizing night time human behaviour as it relates to residual malaria transmission in sub-Saharan Africa: A review of the published literature," *Malaria Journal*, 2019.

[16] H. Ranson and N. Lissenden, "Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control," *Trends in Parasitology*, vol. 32, no. 3, pp. 187–196, 2016.

[17] N. J. White, S. Pukrittayakamee, T. T. Hien, M. A. Faiz, O. A. Mokuolu, and A. M. Dondorp, "Malaria," *The Lancet*, vol. 383, pp. 723–735, 2 2014.

[18] N. J. White, "Plasmodium knowlesi: The Fifth Human Malaria Parasite," *Clinical Infectious Diseases*, vol. 46, pp. 172–173, 1 2008.

[19] WHO, "Malaria entomology and vector control: Guide for participants," tech. rep., WHO, Geneva, 2013.

[20] J. R. Ohm, F. Baldini, P. Barreaux, T. Lefevre, P. A. Lynch, E. Suh, S. A. Whitehead, and M. B. Thomas, "Rethinking the extrinsic incubation period of malaria parasites," *Parasites & Vectors*, vol. 11, no. 1, p. 178, 2018.

[21] M. Balkew, P. Mumba, D. Dengela, G. Yohannes, D. Getachew, S. Yared, S. Chibsa, M. Murphy, K. George, K. Lopez, D. Janies, S. H. Choi, J. Spear, S. R. Irish, and T. E. Carter, "Geographical distribution of Anopheles stephensi in eastern Ethiopia," *Parasites & Vectors*, vol. 13, p. 35, 12 2020.

[22] N. F. Kahamba, F. O. Okumu, M. Jumanne, K. Kifungo, J. O. Odero, F. Baldini, H. M. Ferguson, and L. Nelli, "Geospatial modelling of dry season habitats of the malaria

vector, Anopheles funestus, in south-eastern Tanzania," *Parasites & Vectors*, vol. 17, p. 38, 1 2024.

[23] I. H. Nambunga, H. S. Ngowo, S. A. Mapua, E. E. Hape, B. J. Msugupakulya, D. S. Msaky, N. T. Mhumbira, K. R. Mchwembo, G. Z. Tamayamali, S. V. Mlembe, R. M. Njalambaha, D. W. Lwetoijera, M. F. Finda, N. J. Govella, D. Matoke-Muhia, E. W. Kaindoa, and F. O. Okumu, "Aquatic habitats of the malaria vector Anopheles funestus in rural south-eastern Tanzania," *Malaria Journal*, vol. 19, no. 1, p. 219, 2020.

[24] C. L. Lyons, M. Coetzee, and S. L. Chown, "Stable and fluctuating temperature effects on the development rate and survival of two malaria vectors, Anopheles arabiensis and Anopheles funestus," *Parasites & Vectors*, vol. 6, p. 104, 12 2013.

[25] M. E. Sinka, M. J. Bangs, S. Manguin, Y. Rubio-Palis, T. Chareonviriyaphap, M. Coetzee, C. M. Mbogo, J. Hemingway, A. P. Patil, and W. H. Temperley, "A global map of dominant malaria vectors," *Parasit Vectors*, vol. 5, no. 1, p. 69, 2012.

[26] M. T. Gillies and M. Coetzee, "A Supplement to the Anophelinae of the South of the Sahara (Afrotropical Region).," *Publications of the South African Institute for Medical Research*, vol. 55, pp. 1–143, 1987.

[27] G. F. Killeen, "Characterizing, controlling and eliminating residual malaria transmission," *Malaria Journal*, vol. 13, no. 1, p. 330, 2014.

[28] C. D. Mwalimu, S. Kiware, R. Nshama, Y. Derua, P. Machafuko, P. Gitanya, W. Mwafongo, J. Bernard, B. Emidi, V. Mwingira, R. Malima, V. Githu, B. Masanja, Y. Mlacha, P. Tungu, B. Kabula, E. Sambu, B. Batengana, J. Matowo, N. Govella, P. Chaki, S. Lazaro, N. Serbantez, J. Kitau, S. M. Magesa, and W. N. Kisinza, "Dynamics of malaria vector composition and Plasmodium falciparum infection in mainland Tanzania: 2017–2021 data from the national malaria vector entomological surveillance," *Malaria Journal*, vol. 23, no. 1, p. 29, 2024.

[29] M. K. Faulde, L. M. Rueda, and B. A. Khaireh, "First record of the Asian malaria vector Anopheles stephensi and its possible role in the resurgence of malaria in Djibouti, Horn of Africa," *Acta tropica*, vol. 139, pp. 39–43, 2014.

[30] A. Mnzava, A. C. Monroe, and F. Okumu, "Anopheles stephensi in Africa requires a more integrated response," *Malaria Journal*, vol. 21, no. 1, pp. 1–6, 2022.

[31] E. Ochomo, S. Milanoi, B. Abong'o, B. Onyango, M. Muchoki, D. Omoke, E. Olanga, L. Njoroge, E. O. Juma, J. D. Otieno, D. Matoke-Muhia, L. Kamau, C. Rafferty, J. Gimnig, M. Shieshia, D. Wacira, J. Mwangangi, M. Maia, C. Chege, A. Omar, M. Rono, L. Abel, W. P. O'Meara, A. Obala, C. Mbogo, and L. Kariuki, "Detection of Anopheles stephensi Mosquitoes by Molecular Surveillance, Kenya," *Emerging Infectious Disease journal*, vol. 29, no. 12, p. 2498, 2023.

[32] WHO, *Global technical strategy for malaria 2016-2030*. WHO, 2015.

[33] M. H. Birley and J. D. Charlewood, "Sporozoite rate and malaria prevalence," *Parasitology Today*, vol. 3, pp. 231–232, 8 1987.

[34] T. R. Burkot, R. Farlow, M. Min, E. Espino, A. Mnzava, and T. L. Russell, "A global analysis of National Malaria Control Programme vector surveillance by elimination and control status in 2018," *Malaria Journal*, vol. 18, no. 1, p. 399, 2019.

[35] G. MacDonald, "Epidemiological basis of malaria control.," *Bull. Wld. Hlth. Org*, vol. 15, no. 3-5, pp. 613–626, 1956.

[36] J. C. Beier, G. F. Killeen, and J. Githure, "Short report: Entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa," *Am J Trop Med Hyg*, vol. 61, no. 1, pp. 109–113, 1999.

[37] T. Smith, G. Killeen, C. Lengeler, and M. Tanner, "Relationships between the outcome of Plasmodium falciparum infection and the intensity of transmission in Africa," *The American Journal of Tropical Medicine and Hygiene Am J Trop Med Hyg*, vol. 71, no. 2_suppl, pp. 80–86, 2004.

[38] P. Corran, P. Coleman, E. Riley, and C. Drakeley, "Serology: a robust indicator of malaria transmission intensity?," *Trends in parasitology*, vol. 23, no. 12, pp. 575–582, 2007.

[39] M. F. Finda, A. J. Limwagu, H. S. Ngowo, N. S. Matowo, J. K. Swai, E. Kaindoa, and F. O. Okumu, "Dramatic decreases of malaria transmission intensities in Ifakara, south-eastern Tanzania since early 2000s," *Malaria Journal*, vol. 17, p. 362, 2018.

[40] T. R. Burkot, H. Bugoro, A. Apairamo, R. D. Cooper, D. F. Echeverry, D. Odabasi, N. W. Beebe, V. Makuru, H. Xiao, J. R. Davidson, N. A. Deason, H. Reuben, J. W. Kazura, F. H. Collins, N. F. Lobo, and T. L. Russell, "Spatial-temporal heterogeneity in malaria receptivity is best estimated by vector biting rates in areas nearing elimination," *Parasites & Vectors*, vol. 11, no. 1, p. 606, 2018.

[41] A. E. Kiswewski, A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs, "A global index representing the stability of malaria transmission," *Am J Trop Med Hyg*, vol. 70, pp. 486–498, 2004.

[42] A. Pastusiak, M. R. Reddy, X. Chen, I. Hoyer, J. Dorman, M. E. Gebhardt, G. Carpi, D. E. Norris, J. M. Pipas, and E. K. Jackson, "A metagenomic analysis of the phase 2 Anopheles gambiae 1000 genomes dataset reveals a wide diversity of cobionts associated with field collected mosquitoes," *Communications Biology*, vol. 7, no. 1, p. 667, 2024.

[43] A. D. T. Barrett and T. P. Monath, "Epidemiology and ecology of yellow fever virus," *Advances in virus research*, vol. 61, pp. 291–317, 2003.

[44] B. H. Bird, T. G. Ksiazek, S. T. Nichol, and N. J. MacLachlan, "Rift Valley fever virus," *Journal of the American Veterinary Medical Association*, vol. 234, no. 7, pp. 883–893, 2009.

[45] G. L. Campbell, A. A. Marfin, R. S. Lanciotti, and D. J. Gubler, "West Nile virus," *The Lancet Infectious Diseases*, vol. 2, pp. 519–529, 9 2002.

[46] T. P. Endy and A. Nisalak, "Japanese Encephalitis Virus: Ecology and Epidemiology BT - Japanese Encephalitis and West Nile Viruses," in *Current Topics in Microbiology and Immunology* (J. S. Mackenzie, A. D. T. Barrett, and V. Deubel, eds.), pp. 11–48, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[47] S. B. Halstead, "Dengue Virus–Mosquito Interactions," *Annual Review of Entomology*, vol. 53, pp. 273–291, 12 2007.

[48] L. R. Petersen, D. J. Jamieson, A. M. Powers, and M. A. Honein, "Zika Virus," *New England Journal of Medicine*, vol. 374, pp. 1552–1563, 3 2016.

[49] G. Pialoux, B.-A. Gaüzère, S. Jauréguiberry, and M. Strobel, "Chikungunya, an epidemic arbovirosis," *The Lancet Infectious Diseases*, vol. 7, pp. 319–327, 5 2007.

[50] T. Habtewold, A. A. Sharma, C. A. S. Wyer, E. K. G. Masters, N. Windbichler, and G. K. Christophides, "Plasmodium oocysts respond with dormancy to crowding and nutritional stress," *Scientific Reports*, vol. 11, no. 1, p. 3090, 2021.

[51] H. Hurd, J. C. Hogg, and M. Renshaw, "Interactions between bloodfeeding, fecundity and infection in mosquitoes," *Parasitology Today*, vol. 11, no. 11, pp. 411–416, 1995.

[52] J. D. Charlwood, T. Smith, P. F. Billingsley, W. Takken, E. O. L. Lyimo, and J. Meuwissen, "Survival and infection probabilities of anthropophagic anophelines from an area of high prevalence of Plasmodium falciparum in humans," *Bull. Entomol. Res.*, vol. 87, pp. 445–453, 1997.

[53] T. S. Detinova, "Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria.," *Monograph series. World Health Organization*, 1962.

[54] V. P. Polovodova, "The determination of the physiological age of female Anopheles by the number of gonotrophic cycles completed," *Medskaya. Parazit.*, vol. 18, pp. 352–355, 1949.

[55] V. Rao, "On Gonotrophic Discordance among certain Indian Anopheles.," *Indian Journal of Malariology*, vol. 1, pp. 43–50, 1947.

[56] 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*, vol. 103, pp. 18060–18065, 11 2006.

[57] I. Iovinella, B. Caputo, E. Michelucci, F. R. Dani, and A. della Torre, "Candidate biomarkers for mosquito age-grading identified by label-free quantitative analysis of protein expression in Aedes albopictus females," *Journal of Proteomics*, vol. 128, pp. 272–279, 2015.

[58] L. Gray, B. C. Asay, B. Hephaestus, R. McCabe, G. Pugh, E. D. Markle, T. S. Churcher, and B. D. Foy, "Back to the Future: Quantifying Wing Wear as a Method to Measure Mosquito Age," *The American Journal of Tropical Medicine and Hygiene*, vol. 107, no. 3, pp. 689–700, 2022.

[59] M. Gonzalez-Jimenez, S. A. Babayan, P. Khazaeli, M. Doyle, F. Walton, E. Reedy, T. Glew, M. Viana, L. Ranford-Cartwright, A. Niang, D. Siria, F. O. Okumu, A. Diabate, 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," *Wellcome Open Res*, vol. 4:76, 2019.

[60] 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 sl using near-infrared spectroscopy," *American Journal of Tropical Medicine and Hygiene*, vol. 81, no. 4, p. 622, 2009.

[61] D. J. Siria, R. Sanou, J. Mitton, E. P. Mwanga, A. Niang, I. Sare, P. C. D. Johnson, G. M. 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 age-grading and species identification of natural mosquitoes for malaria surveillance," *Nature Communications*, vol. 13, no. 1, p. 1501, 2022.

[62] L. F. Chaves, L. C. Harrington, C. L. Keogh, A. M. Nguyen, and U. D. Kitron, "Blood feeding patterns of mosquitoes: Random or structured?," *Frontiers in Zoology*, vol. 7, no. 1, p. 3, 2010.

[63] A. N. Clements, *The Biology of Mosquitoes 1: Development, Nutrition and Reproduction*, vol. 1. London: Chapman & Hall, 1992.

[64] H. Iwashita, G. O. Dida, G. O. Sonye, T. Sunahara, K. Futami, S. M. Njenga, L. F. Chaves, and N. Minakawa, "Push by a net, pull by a cow: can zooprophylaxis enhance the impact of insecticide treated bed nets on malaria control?," *Parasites & Vectors*, vol. 7, no. 1, p. 52, 2014.

[65] B. Donnelly, L. Berrang-Ford, N. A. Ross, and P. Michel, "A systematic, realist review of zooprophylaxis for malaria control," *Malaria Journal*, vol. 14, no. 1, p. 313, 2015.

[66] H. Hasyim, M. Dhimal, J. Bauer, D. Montag, D. A. Groneberg, U. Kuch, and R. Müller, "Does livestock protect from malaria or facilitate malaria prevalence? A cross-sectional study in endemic rural areas of Indonesia," *Malaria Journal*, vol. 17, no. 1, p. 302, 2018.

[67] M. Bouma and M. Rowland, "Failure of passive zooprophylaxis: cattle ownership in Pakistan is associated with a higher malaria prevalence," *Trans. Roy. Soc. Trop. Med. Hyg.*, vol. 89, pp. 351–353, 1995.

[68] J. C. Beier, P. V. Perkins, R. A. Wirtz, J. Koros, D. Diggs, T. P. Gargan, and D. K. Koech, "Bloodmeal identification by direct enzyme-linked immunosorbent assay (ELISA), tested on Anopheles (Diptera: Culicidae) in Kenya.," *Journal of medical entomology*, vol. 25, no. 1, pp. 9–16, 1988.

[69] R. J. Kent and D. E. Norris, "Identification of mammalian blood meals in mosquitoes by a multiplexed polymerase chain reaction targeting cytochrome B," *The American Journal of Tropical Medicine and Hygiene*, vol. 73, pp. 336–342, 8 2005.

[70] Y. Hatefi, "The Mitochondrial Electron Transport and Oxidative Phosphorylation System," *Annual Review of Biochemistry*, vol. 54, no. 1, pp. 1015–1069, 1985.

[71] D. M. Irwin, T. D. Kocher, and A. C. Wilson, "Evolution of the cytochromeb gene of mammals," *Journal of Molecular Evolution*, vol. 32, no. 2, pp. 128–144, 1991.

[72] L. E. Hugo, P. E. Cook, P. H. Johnson, L. P. Rapley, B. H. Kay, P. A. Ryan, S. A. Ritchie, and S. L. O'Neill, "Field Validation of a Transcriptional Assay for the Prediction of Age of Uncaged Aedes aegypti Mosquitoes in Northern Australia," *PLOS Neglected Tropical Diseases*, vol. 4, p. e608, 2 2010.

[73] Z. T. Nagy, "A hands-on overview of tissue preservation methods for molecular genetic analyses," *Organisms Diversity & Evolution*, vol. 10, no. 1, pp. 91–105, 2010.

[74] E. Chow, R. A. wirtz, and T. W. Scott, "Identification of blood meals in Aedesaegypti by antibody sandwich enzyme-linked immunosorbent assay," *Journal of the American Mosquito Control Association*, 1993.

[75] J. C. Beier, P. V. Perkins, J. K. Koros, F. K. Onyango, T. P. Gargan, R. A. Wirtz, D. K. Koech, and C. R. Roberts, "Malaria sporozoite detection by dissection and ELISA to assess infectivity of afrotropical Anopheles (Diptera: Culicidae).," *Journal of medical entomology*, vol. 27, no. 3, pp. 377–384, 1990.

[76] H. Aonuma, M. Suzuki, H. Iseki, N. Perera, B. Nelson, I. Igarashi, T. Yagi, H. Kanuka, and S. Fukumoto, "Rapid identification of Plasmodium-carrying mosquitoes

using loop-mediated isothermal amplification," *Biochemical and Biophysical Research Communications*, vol. 376, no. 4, pp. 671–676, 2008.

[77] T. Notomi, H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino, and T. Hase, "Loop-mediated isothermal amplification of DNA," *Nucleic Acids Research*, vol. 28, pp. e63–e63, 6 2000.

[78] C. Bass, D. Nikou, A. M. Blagborough, J. Vontas, R. E. Sinden, M. S. Williamson, and L. M. Field, "PCR-based detection of Plasmodium in Anopheles mosquitoes: a comparison of a new high-throughput assay with existing methods," *Malaria Journal*, vol. 7, no. 1, p. 177, 2008.

[79] V. Boonsaeng, S. Panyim, P. Wilairat, and A. Tassanakajon, "Polymerase chain reaction detection of Plasmodium falciparum in mosquitoes," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 87, no. 3, pp. 273–275, 1993.

[80] M. S. Beier, I. K. Schwartz, J. C. Beier, P. V. Perkins, F. Onyango, J. K. Koros, G. H. Campbell, P. M. Andrysiak, and A. D. Brandling-Bennett, "Identification of Malaria Species by Elisa in Sporozoite and Oocyst Infected Anopheles from Western Kenya," *The American Journal of Tropical Medicine and Hygiene*, vol. 39, no. 4, pp. 323–327, 1988.

[81] T. R. Burkot, J. L. Williams, and I. Schneider, "Identification of Plasmodium falciparum-infected mosquitoes by a double antibody enzyme-linked immunosorbent assay," *American Journal of Tropical Medicine and Hygiene*, vol. 33, pp. 783–788, 1984.

[82] T. R. Burkot, P. M. Graves, J. A. Cattani, R. W. Wirtz, and F. D. Gibson, "The efficiency of sporozoite transmission in the human malarias, Plasmodium falciparum and P. vivax," *Bull. Wld. Hlth. Org.*, vol. 65, no. 3, pp. 375–380, 1987.

[83] F. H. Collins, P. M. Procell, G. H. Campbell, and W. E. Collins, "Monoclonal Antibody-Based Enzyme-Linked Immunosorbent Assay (Elisa) for Detection of Plasmodium malariae Sporozoites in Mosquitoes," *The American Journal of Tropical Medicine and Hygiene*, vol. 38, no. 2, pp. 283–288, 1988.

[84] K. Bashar, N. Tuno, T. U. Ahmed, and A. J. Howlader, "False positivity of circumsporozoite protein (CSP)-ELISA in zoophilic anophelines in Bangladesh," *Acta Tropica*, vol. 125, no. 2, pp. 220–225, 2013.

[85] L. Durnez, W. Van Bortel, L. Denis, P. Roelants, A. A. Veracx, H. D. Trung, T. Sochantha, and M. Coosemans, "False positive circumsporozoite protein ELISA: a challenge for the estimation of the entomological inoculation rate of malaria and for vector incrimination," *Malaria Journal*, vol. 10, p. 195, 2011.

[86] A. Mukhwana, O. Shorinola, D. Ndlovu, and J. Osaso, "Slow, difficult and expensive: How the lab supplies market is crippling African science."

[87] H. Büning-Pfaue, "Analysis of water in food by near infrared spectroscopy," *Food Chemistry*, vol. 82, no. 1, pp. 107–115, 2003.

[88] G. Qian and Z. Y. Wang, "Near-Infrared Organic Compounds and Emerging Applications," *Chemistry – An Asian Journal*, vol. 5, pp. 1006–1029, 5 2010.

[89] S. Fuentes, E. Tongson, R. R. Unnithan, and C. Gonzalez Viejo, "Early Detection of Aphid Infestation and Insect-Plant Interaction Assessment in Wheat Using a Low-Cost Electronic Nose (E-Nose), Near-Infrared Spectroscopy and Machine Learning Modeling," 2021.

[90] 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," *Parasites and Vectors*, vol. 6, no. 1, p. 298, 2013.

[91] M. Jamrógiewicz, "Application of the near-infrared spectroscopy in the pharmaceutical technology," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 66, pp. 1–10, 2012.

[92] D. F. Da, R. McCabe, B. M. Somé, P. M. Esperança, K. A. Sala, J. Blight, A. M. Blagborough, F. Dowell, S. R. Yerbanga, T. Lefèvre, K. Mouline, R. K. Dabiré, and T. S. Churcher, "Detection of Plasmodium falciparum in laboratory-reared and naturally infected wild mosquitoes using near-infrared spectroscopy," *Scientific Reports*, vol. 11, no. 1, p. 10289, 2021.

[93] M. F. Maia, M. Kapulu, M. Muthui, M. G. Wagah, H. M. Ferguson, F. E. Dowell, F. Baldini, and L.-R. L. R. Cartwright, "Detection of Plasmodium falciparum infected Anopheles gambiae using near-infrared spectroscopy," *Malaria Journal*, vol. 18, no. 1, p. 85, 2019.

[94] 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, no. 1, 2018.

[95] 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, 2016.

[96] L. Genzel, F. Kremer, A. Poglitsch, and G. Bechtold, "Millimeter-wave and Far-infrared Spectroscopy on Biological Macromolecules BT - Coherent Excitations in Biological Systems," in *Coherent Excitations in Biological Systems* (H. Fröhlich and F. Kremer, eds.), (Berlin, Heidelberg), pp. 58–70, Springer Berlin Heidelberg, 1983.

[97] G. Gauglitz and D. S. Moore, *Handbook of Spectroscopy: Second, Enlarged Edition*. Wiley, 2nd edition ed., 2014.

[98] Metrohm, *NIR Spectroscopy : A guide to near-infrared spectroscopic analysis of industrial manufacturing processes*. Herisau: Metrohm AG, 2013.

[99] A. Paudel, D. Raijada, and J. Rantanen, "Raman spectroscopy in pharmaceutical product design," *Advanced Drug Delivery Reviews*, vol. 89, pp. 3–20, 2015.

[100] T. Vankeirsbilck, A. Vercauteren, W. Baeyens, G. Van der Weken, F. Verpoort, G. Vergote, and J. P. Remon, "Applications of Raman spectroscopy in pharmaceutical analysis," *TrAC Trends in Analytical Chemistry*, vol. 21, no. 12, pp. 869–877, 2002.

[101] D. R. Neuville, D. de Ligny, and G. S. Henderson, "Advances in Raman Spectroscopy Applied to Earth and Material Sciences," *Reviews in Mineralogy and Geochemistry*, vol. 78, pp. 509–541, 1 2014.

[102] C. Muehlethaler, M. Leona, and J. R. Lombardi, "Review of Surface Enhanced Raman Scattering Applications in Forensic Science," *Analytical Chemistry*, vol. 88, pp. 152–169, 1 2016.

[103] Z. Gao, L. C. Harrington, W. Zhu, L. M. Barrientos, C. Alfonso-Parra, F. W. Avila, J. M. Clark, and L. He, "Accurate age-grading of field-aged mosquitoes reared under ambient conditions using surface-enhanced Raman spectroscopy and artificial neural networks," *Journal of Medical Entomology*, vol. 60, pp. 917–923, 9 2023.

[104] 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, 2 2022.

[105] B. M. Tissue, "Ultraviolet and Visible Absorption Spectroscopy," in *Characterization of Materials*, Wiley, 10 2002.

[106] J. R. Lakowicz, "Instrumentation for Fluorescence Spectroscopy," in *Principles of Fluorescence Spectroscopy*, pp. 27–61, Boston, MA: Springer US, 2006.

[107] J. A. Adegoke, A. De Paoli, I. O. Afara, K. Kochan, D. J. Creek, P. Heraud, and B. R. Wood, "Ultraviolet/Visible and Near-Infrared Dual Spectroscopic Method for Detection and Quantification of Low-Level Malaria Parasitemia in Whole Blood," *Analytical Chemistry*, vol. 93, pp. 13302–13310, 10 2021.

[108] S. Attal, R. Thiruvengadathan, and O. Regev, "Determination of the Concentration of Single-Walled Carbon Nanotubes in Aqueous Dispersions Using UV−Visible Absorption Spectroscopy," *Analytical Chemistry*, vol. 78, pp. 8098–8104, 12 2006.

[109] Y. M. Serebrennikova, J. Patel, and L. H. Garcia-Rubio, "Interpretation of the ultraviolet-visible spectra of malaria parasite Plasmodium falciparum," *Applied Optics*, vol. 49, no. 2, pp. 180–188, 2010.

[110] Z. Shi, C. W. K. Chow, R. Fabris, J. Liu, and B. Jin, "Applications of Online UV-Vis Spectrophotometer for Drinking Water Quality Monitoring and Process Control: A Review," 2022.

[111] T. Shi, Y. Chen, Y. Liu, and G. Wu, "Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals," *Journal of Hazardous Materials*, vol. 265, pp. 166–176, 2014.

[112] S. L. Upstone, "Ultraviolet/Visible Light Absorption Spectrophotometry in Clinical Chemistry Update based on the original article by Stephen L. Upstone, Encyclopedia of Analytical Chemistry, © 2000, John Wiley & Sons, Ltd.," in *Encyclopedia of Analytical Chemistry*, Wiley, 3 2013.

[113] A. Khoshmanesh, M. W. A. Dixon, S. Kenny, L. Tilley, D. McNaughton, and B. R. Wood, "Detection and quantification of early-stage malaria parasites in laboratory infected erythrocytes by attenuated total reflectance infrared spectroscopy and multivariate analysis," *Analytical Chemistry*, vol. 86, no. 9, p. 4379–4386, 2014.

[114] 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, 5 2019.

[115] E. P. Mwanga, E. G. Minja, E. Mrimi, M. González-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, 2019.

[116] 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, no. 1, pp. 157–163, 2011.

[117] O. R. Wood, S. Hanrahan, M. Coetzee, L. L. Koekemoer, and B. D. Brooke, "Cuticle thickening associated with pyrethroid resistance in the major malaria vector Anopheles funestus," *Parasites & Vectors*, vol. 3, no. 1, p. 67, 2010.

[118] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. Boston: O'Reilly Media, Inc., first edit ed., 2017.

[119] A. Panesar, *Machine Learning and AI for Healthcare*. Berkeley: Apress, 2 ed., 2021.

[120] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 1 ed., 2006.

[121] T. J. Cleophas and A. H. Zwinderman, *Machine Learning in Medicine – A Complete Overview*. Cham: Springer International Publishing, 2 ed., 2020.

[122] O. V. Prezhdo, "Advancing Physical Chemistry with Machine Learning," *The Journal of Physical Chemistry Letters*, vol. 11, pp. 9656–9658, 11 2020.

[123] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2 ed., 2009.

[124] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, no. 7, pp. 641–642, 2017.

[125] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, 1987.

[126] 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.

[127] L. Djamouko-Djonkam, D. L. Nkahe, E. Kopya, A. Talipouo, C. S. Ngadjeu, P. Doumbe-Belisse, R. Bamou, P. Awono-Ambene, T. Tchuinkam, C. S. Wondji, and C. Antonio-Nkondjio, "Implication of Anopheles funestus in malaria transmission in the city of Yaoundé, Cameroon TT - Implication d'Anopheles funestus dans la transmission du paludisme dans la ville de Yaoundé au Cameroun," *Parasite (Paris, France)*, vol. 27, p. 10, 2020.

[128] E. W. Kaindoa, N. S. Matowo, H. S. Ngowo, G. Mkandawile, A. Mmbando, M. Finda, and F. O. Okumu, "Interventions that effectively target Anopheles funestus mosquitoes could significantly improve control of persistent malaria transmission in south–eastern Tanzania," *PloS one*, vol. 12, no. 5, p. e0177807, 2017.

[129] D. W. Lwetoijera, C. Harris, S. S. Kiware, S. Dongus, G. J. Devine, P. J. McCall, and S. Majambere, "Increasing role of Anopheles funestus and Anopheles arabiensis in malaria transmission in the Kilombero Valley, Tanzania," *Malaria Journal*, vol. 13, no. 1, 2014.

[130] S. A. Mapua, E. E. Hape, J. Kihonda, H. Bwanary, K. Kifungo, M. Kilalangongono, E. W. Kaindoa, H. S. Ngowo, and F. O. Okumu, "Persistently high proportions of plasmodium-infected Anopheles funestus mosquitoes in two villages in the

Kilombero valley, South-Eastern Tanzania," *Parasite Epidemiology and Control*, vol. 18, p. e00264, 2022.

[131] N. S. Matowo, M. A. Kulkarni, L. A. Messenger, M. Jumanne, J. Martin, E. Mallya, E. Lukole, J. F. Mosha, O. Moshi, B. Shirima, R. Kaaya, M. Rowland, A. Manjurano, F. W. Mosha, and N. Protopopoff, "Differential impact of dual-active ingredient long-lasting insecticidal nets on primary malaria vectors: a secondary analysis of a 3-year, single-blind, cluster-randomised controlled trial in rural Tanzania," *The Lancet Planetary Health*, vol. 7, pp. e370–e380, 5 2023.

[132] B. J. Msugupakulya, N. H. Urio, M. Jumanne, H. S. Ngowo, P. Selvaraj, F. O. Okumu, and A. L. Wilson, "Changes in contributions of different Anopheles vector species to malaria transmission in east and southern Africa from 2000 to 2022," *Parasites & Vectors*, vol. 16, no. 1, p. 408, 2023.

[133] 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 & Vectors*, vol. 11, no. 1, p. 577, 2018.

[134] J. K. Swai, A. S. Mmbando, H. S. Ngowo, O. G. Odufuwa, M. F. Finda, W. Mponzi, A. P. Nyoni, D. Kazimbaya, A. J. Limwagu, R. M. Njalambaha, S. Abbasi, S. J. Moore, J. Schellenberg, L. M. Lorenz, and F. O. Okumu, "Protecting migratory farmers in rural Tanzania using eave ribbons treated with the spatial mosquito repellent, transfluthrin," *Malaria Journal*, vol. 18, no. 1, p. 414, 2019.

[135] 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, no. 1, p. 408, 2020.

[136] W. Takken and N. O. Verhulst, "Host Preferences of Blood-Feeding Mosquitoes," *Annual Review of Entomology*, vol. 58, no. 1, pp. 433–453, 2013.

[137] WHO, "World Malaria Report 2021," tech. rep., WHO, Geneva, 2021.

[138] P. M. Esperança, D. F. Da, B. Lambert, R. K. Dabiré, and T. S. Churcher, "Functional data analysis techniques to improve the generalizability of near-infrared spectral data for monitoring mosquito populations," *bioRxiv*, p. 2020.04.28.058495, 1 2020.

[139] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*, (Berlin, Heidelberg), pp. 583–588, Springer, 1997.

[140] L. J. P. Van Der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, 2008.

[141] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, V. Dubourg, F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, F. Pedregosa, and R. Weiss, "Scikit-learn : Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12(85), p. 2825−2830, 2011.

[142] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *30th International Conference on Machine Learning, ICML 2013*, 2013.

[143] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *International Joint Conference of Artificial Intelligence*, 1995.

[144] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 2016.

[145] F. Chollet, "Keras: The Python Deep Learning library," *Keras.Io*, 2015.

[146] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, vol. 53, pp. 217–288, 1 2011.

[147] 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, no. 1, 2018.

[148] B. Hanczar, V. Bourgeais, and F. Zehraoui, "Assessment of deep learning and transfer learning for cancer prediction based on gene expression data," *BMC Bioinformatics*, vol. 23, no. 1, p. 262, 2022.

[149] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation Regularization: A General Framework for Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[150] S. Si, D. Tao, and B. Geng, "Bregman Divergence-Based Regularization for Transfer Subspace Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 929–942, 2010.

[151] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain Adaptation via Transfer Component Analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[152] P. Mignone, G. Pio, S. Džeroski, and M. Ceci, "Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks," *Scientific Reports*, vol. 10, no. 1, p. 22295, 2020.

[153] G. Pio, P. Mignone, G. Magazzù, G. Zampieri, M. Ceci, and C. Angione, "Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction," *Bioinformatics*, vol. 38, pp. 487–493, 1 2022.

[154] A. Mbengue, S. Bhattacharjee, T. Pandharkar, H. Liu, G. Estiu, R. V. Stahelin, S. S. Rizk, D. L. Njimoh, Y. Ryan, K. Chotivanich, C. Nguon, M. Ghorbal, J.-J. Lopez-Rubio, M. Pfrender, S. Emrich, N. Mohandas, A. M. Dondorp, O. Wiest, and K. Haldar, "A molecular mechanism of artemisinin resistance in Plasmodium falciparum malaria," *Nature*, vol. 520, no. 7549, pp. 683–687, 2015.

[155] G. Siddiqui, A. Srivastava, A. S. Russell, and D. J. Creek, "Multi-omics Based Identification of Specific Biochemical Changes Associated With PfKelch13-Mutant Artemisinin-Resistant Plasmodium falciparum," *The Journal of Infectious Diseases*, vol. 215, pp. 1435–1444, 5 2017.

[156] E. A. Winzeler and M. J. Manary, "Drug resistance genomics of the antimalarial drug artemisinin," *Genome Biology*, vol. 15, no. 11, p. 544, 2014.

[157] C. Sokhna, M. O. Ndiath, and C. Rogier, "The changes in mosquito vector behaviour and the emerging resistance to insecticides will challenge the decline of malaria," *Clinical Microbiology and Infection*, vol. 19, no. 10, pp. 902–907, 2013.

[158] M. Weill, G. Lutfalla, K. Mogensen, F. Chandre, A. Berthomieu, C. Berticat, N. Pasteur, A. Philips, P. Fort, and M. Raymond, "Insecticide resistance in mosquito vectors," *Nature*, vol. 423, no. 6936, pp. 136–137, 2003.

[159] B. B. Agaba, A. Yeka, S. Nsobya, E. Arinaitwe, J. Nankabirwa, J. Opigo, P. Mbaka, C. S. Lim, J. N. Kalyango, C. Karamagi, and M. R. Kamya, "Systematic review of the status of pfhrp2 and pfhrp3 gene deletion, approaches and methods used for its estimation and reporting in Plasmodium falciparum populations in Africa: review of published studies 2010–2019," *Malaria Journal*, vol. 18, no. 1, p. 355, 2019.

[160] P. Berzosa, V. González, L. Taravillo, A. Mayor, M. Romay-Barja, L. García, P. Ncogo, M. Riloha, and A. Benito, "First evidence of the deletion in the pfhrp2 and pfhrp3 genes in Plasmodium falciparum from Equatorial Guinea," *Malaria Journal*, vol. 19, no. 1, p. 99, 2020.

[161] A. B. Bosco, K. Anderson, K. Gresty, C. Prosser, D. Smith, J. I. Nankabirwa, S. Nsobya, A. Yeka, J. Opigo, S. Gonahasa, R. Namubiru, E. Arinaitwe, P. Mbaka, J. Kissa, S. Won, B. Lee, C. S. Lim, C. Karamagi, J. Cunningham, J. K. Nakayaga, M. R. Kamya, and Q. Cheng, "Molecular surveillance reveals the presence of pfhrp2 and pfhrp3 gene

deletions in Plasmodium falciparum parasite populations in Uganda, 2017–2019," *Malaria Journal*, vol. 19, no. 1, p. 300, 2020.

[162] R. Funwei, D. Nderu, C. N. Nguetse, B. N. Thomas, C. O. Falade, T. P. Velavan, and O. Ojurongbe, "Molecular surveillance of pfhrp2 and pfhrp3 genes deletion in Plasmodium falciparum isolates and the implications for rapid diagnostic tests in Nigeria," *Acta Tropica*, vol. 196, pp. 121–125, 2019.

[163] R. Thomson, K. B. Beshir, J. Cunningham, F. Baiden, J. Bharmal, K. J. Bruxvoort, C. Maiteki-Sebuguzi, S. Owusu-Agyei, S. G. Staedke, and H. Hopkins, "pfhrp2 and pfhrp3 Gene Deletions That Affect Malaria Rapid Diagnostic Tests for Plasmodium falciparum: Analysis of Archived Blood Samples From 3 African Countries," *The Journal of Infectious Diseases*, vol. 220, pp. 1444–1452, 9 2019.

[164] A. S. Parpia, M. L. Ndeffo-Mbah, N. S. Wenzel, and A. P. Galvani, "Effects of response to 2014–2015 Ebola outbreak on deaths from malaria, HIV/AIDS, and tuberculosis, West Africa," *Emerging infectious diseases*, vol. 22, no. 3, p. 433, 2016.

[165] E. Sherrard-Smith, A. B. Hogan, A. Hamlet, O. J. Watson, C. Whittaker, P. Winskill, F. Ali, A. B. Mohammad, P. Uhomoibhi, and I. Maikore, "The potential public health consequences of COVID-19 on malaria in Africa," *Nature medicine*, vol. 26, no. 9, pp. 1411–1416, 2020.

[166] A. J. Limwagu, E. W. Kaindoa, H. S. Ngowo, E. Hape, M. Finda, G. Mkandawile, J. Kihonda, K. Kifungo, R. M. Njalambaha, D. Matoke-Muhia, and F. O. Okumu, "Using a miniaturized double-net trap (DN-Mini) to assess relationships between indoor–outdoor biting preferences and physiological ages of two malaria vectors, Anopheles arabiensis and Anopheles funestus," *Malaria Journal*, vol. 18, no. 1, p. 282, 2019.

[167] J. T. Midega, C. M. Mbogo, H. Mwambi, M. D. Wilson, G. Ojwang, J. M. Mwangangi, J. G. Nzovu, J. I. Githure, G. Yan, and J. C. Beier, "Estimating Dispersal and Survival of <I>Anopheles gambiae</I> and <I>Anopheles funestus</I> Along the Kenyan Coast by Using Mark–Release–Recapture Methods," *Journal of Medical Entomology*, 2007.

[168] M. Coetzee and L. L. Koekemoer, "Molecular systematics and insecticide resistance in the major African malaria vector Anopheles funestus," *Annual review of entomology*, vol. 58, pp. 393–412, 2013.

[169] J. D. Charlwood, E. V. E. Tomás, A. K. Andegiorgish, S. Mihreteab, and C. LeClair, "'We like it wet': a comparison between dissection techniques for the assessment of parity in Anopheles arabiensis and determination of sac stage in mosquitoes alive or dead on collection," *PeerJ*, vol. 6, p. e5155, 2018.

[170] J. Derek Charlwood, S. Nenhep, S. Sovannaroth, J. C. Morgan, J. Hemingway, N. Chitnis, and O. J. T. Briët, "'Nature or nurture': survival rate, oviposition interval, and possible gonotrophic discordance among South East Asian anophelines," *Malaria Journal*, vol. 15, no. 1, p. 356, 2016.

[171] 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.," *Bulletin of the Society for Vector Ecology*, vol. 15, no. 2, pp. 129–139, 1990.

[172] 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.

[173] 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, 10 2010.

[174] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis*. CRC press, 2008.

[175] O. T. W. 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 & Vectors*, vol. 13, no. 1, p. 160, 2020.

[176] 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.

[177] 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, p. e90657, 3 2014.

[178] E. P. Mwanga, D. J. Siria, J. Mitton, I. H. Mshani, M. González-Jiménez, P. Selvaraj, K. Wynne, F. Baldini, F. O. Okumu, and S. A. Babayan, "Using transfer learning and dimensionality reduction techniques to improve generalisability of machine-learning predictions of mosquito ages from mid-infrared spectra," *BMC Bioinformatics*, vol. 24, no. 1, p. 11, 2023.

[179] M. Coetzee, "Key to the females of Afrotropical Anopheles mosquitoes (Diptera: Culicidae)," *Malaria Journal*, vol. 19, no. 1, p. 70, 2020.

[180] L. L. Koekemoer, L. Kamau, R. H. Hunt, and M. Coetzee, "A cocktail polymerase chain reaction assay to identify members of the Anopheles funestus (Diptera: Culicidae) group," *Am J Trop Med Hyg*, vol. 66, no. 6, pp. 804–811, 2002.

[181] S. Nitish, H. Geoffrey, K. Alex, S. Ilya, and S. Ruslan, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 2014.

[182] L. Prechelt, "Early stopping - But when?," *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2012.

[183] I. Dia, M. W. Guelbeogo, and D. Ayala, "Advances and Perspectives in the Study of the Malaria Mosquito Anopheles funestus," in *Anopheles mosquitoes - New insights into malaria vectors* (S. Manguin, ed.), p. Ch. 7, Rijeka: IntechOpen, 2013.

[184] WHO, "Malaria surveillance, monitoring & evaluation: a reference manual.," 2018.

[185] I. Tirados, C. Costantini, G. Gibson, and S. J. Torr, "Blood feeding behaviour of the malarial mosquito Anopheles arabiensis: implications for vector control," *Med Vet Entomol*, vol. 20, no. 4, pp. 425–437, 2006.

[186] R. E. Tedrow, T. Rakotomanga, T. Nepomichene, R. E. Howes, J. Ratovonjato, A. C. Ratsimbasoa, G. J. Svenson, and P. A. Zimmerman, "Anopheles mosquito surveillance in Madagascar reveals multiple blood feeding behavior and Plasmodium infection," *PLOS Neglected Tropical Diseases*, vol. 13, p. e0007176, 7 2019.

[187] J. J. Lemasson, D. Fontenille, L. Lochouarn, I. Dia, F. Simard, K. Ba, A. Diop, M. Diatta, and J. F. Molez, "Comparison of behavior and vector efficiency of Anopheles gambiae and An. arabiensis (Diptera:Culicidae) in Barkedji, a Sahelian area of Senegal," *J Med Entomol*, vol. 34, no. 4, pp. 396–403, 1997.

[188] N. Moiroux, M. B. Gomez, C. C. Pennetier, E. Elanga, A. DjÃ¨nontin, F. Chandre, I. Djègbé, H. H. Guis, V. Corbel, A. Djènontin, F. Chandre, I. Djègbé, H. H. Guis, and V. Corbel, "Changes in Anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in Benin," *Journal of Infectious Diseases*, vol. 206, no. 10, pp. 1622–1629, 2012.

[189] S. Omondi, J. Kosgei, G. Musula, M. Muchoki, B. Abong'o, S. Agumba, C. Ogwang, D. McDermott, M. Donnelly, S. Staedke, J. Schultz, J. Gutman, J. Gimnig, and E. Ochomo, "Late morning biting behaviour of Anopheles funestus is a risk factor for transmission in schools in Siaya, western Kenya," 2023.

[190] F. C. Meza, K. S. Kreppel, D. F. Maliti, A. T. Mlwale, N. Mirzai, G. F. Killeen, H. M. Ferguson, and N. J. Govella, "Mosquito electrocuting traps for directly measuring biting rates and host-preferences of Anopheles arabiensis and Anopheles funestus outdoors," *Malaria Journal*, vol. 18, no. 1, p. 83, 2019.

[191] A. Smith, "The attractiveness of an adult and child to A. gambiae.," *East African medical journal*, vol. 33, no. 10, 1956.

[192] J. D. Charlwood and P. M. Graves, "The effect of permethrin-impregnated bednets on a population of Anopheles farauti in coastal Papua New Guinea," *Med Vet Entomol*, vol. 1, pp. 319–327, 1987.

[193] C. N. M. Mbogo, N. M. Baya, A. V. O. Ofulla, J. I. Githure, and R. W. Snow, "The impact of permethrin-impregnated bednets on malaria vectors of the Kenyan coast," *Medical and Veterinary Entomology*, vol. 10, pp. 251–259, 1996.

[194] L. A. Gomes, R. Duarte, D. C. Lima, B. S. Diniz, M. L. Serrão, and N. Labarthe, "Comparison between precipitin and ELISA tests in the bloodmeal detection of Aedes aegypti (Linnaeus) and Aedes fluviatilis (Lutz) mosquitoes experimentally fed on feline, canine and human hosts," 2001.

[195] T. C. Thiemann, A. C. Brault, H. B. Ernest, and W. K. Reisen, "Development of a high-throughput microsphere-based molecular assay to identify 15 common bloodmeal hosts of Culex mosquitoes," *Molecular ecology resources*, vol. 12, no. 2, pp. 238–246, 2012.

[196] J. Ansell, J.-T. Hu, S. C. Gilbert, K. A. Hamilton, A. V. S. Hill, and S. W. Lindsay, "Improved method for distinguishing the human source of mosquito blood meals between close family members," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 94, no. 5, pp. 572–574, 2000.

[197] T. R. Burkot and G. R. DeFoliart, "Bloodmeal sources of Aedes triseriatus and Aedes vexans in a southern Wisconsin forest endemic for La Crosse encephalitis virus ," 1982.

[198] G. Edrissian and A. Hafizi, "Application of enzyme-linked immunosorbent assay (ELISA) to identification of Anopheles mosquito bloodmeals," *Transactions of The Royal Society of Tropical Medicine and Hygiene*, vol. 76, pp. 54–56, 1 1982.

[199] C. Kimpton, A. Walton, and P. Gill, "A further tetranucleotide repeat polymorphism in the vWF gene," *Human Molecular Genetics*, vol. 1, p. 287, 7 1992.

[200] M. H. Polymeropoulos, H. Xiao, D. S. Rath, and C. R. Merrill, "Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH)," 1991.

[201] R. W. Mukabana, W. Takken, P. Seda, G. F. Killeen, W. A. Hawley, and B. G. J. Knols, "Extent of digestion affects the success of amplifying human DNA isolated from blood meals of Anopheles gambiae (Diptera: Culicidae)," *Bulletin of Entomological Research*, vol. 92, no. 3, pp. 233–239, 2002.

[202] M. T. Gillies and M. Coetzee, *A supplement to the Anophelinae of Africa South of the Sahara (Afrotropical region)*. Johannesburg: South African Medical Research Institute, 1987.

[203] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 9 2017.

[204] I. H. Mshani, D. J. Siria, E. P. Mwanga, B. B. D. Sow, R. Sanou, M. Opiyo, M. T. Sikulu-Lord, H. M. Ferguson, A. Diabate, K. Wynne, M. González-Jiménez, F. Baldini, S. A. Babayan, and F. Okumu, "Key considerations, target product profiles, and research gaps in the application of infrared spectroscopy and artificial intelligence for malaria surveillance and diagnosis," *Malaria Journal*, vol. 22, no. 1, p. 346, 2023.

[205] A. K. Githeko, M. W. Service, C. M. Mbogo, F. K. Atieli, and F. O. Juma, "Origin of blood meals in indoor and outdoor resting malaria vectors in western Kenya," *Acta tropica*, vol. 58, pp. 307–316, 1994.

[206] G. C. Katusi, M. R. G. Hermy, S. M. Makayula, R. Ignell, N. J. Govella, S. R. Hill, and L. L. Mnyone, "Seasonal variation in abundance and blood meal sources of primary and secondary malaria vectors within Kilombero Valley, Southern Tanzania," *Parasites & Vectors*, vol. 15, no. 1, p. 479, 2022.

[207] C. Bogh, S. E. Clarke, G. E. L. Walraven, and S. W. Lindsay, "Zooprophylaxis,artefact or reality? A paired-cohort study of the effect of passive zooprophlaxis in malaria in The Gambia," *Trans R Soc Trop Med Hyg*, vol. 96, pp. 593–596, 2002.

[208] C. Bøgh, S. E. Clarke, M. Pinder, F. Sanyang, and S. W. Lindsay, "Effect of Passive Zooprophylaxis on Malaria Transmission in the Gambia," *Journal of Medical Entomology*, vol. 38, pp. 822–828, 11 2001.

[209] A. Saul, "Zooprophylaxis or zoopotentiation: the outcome of introducing animals on vector transmission is highly dependent on the mosquito mortality while searching," *Malar J*, vol. 2, no. 1, pp. 1–18, 2003.

[210] T. Sota and M. Mogi, "Effectiveness of zooprophylaxis in malaria control: a theoretical inquiry with a model for mosquito populations with two bloodmeal hosts," *Med. Vet. Entomol.*, vol. 3, pp. 337–345, 1989.

[211] Y. A. Derua, M. Alifrangis, S. M. Magesa, W. N. Kisinza, and P. E. Simonsen, "Sibling species of the Anopheles funestus group, and their infection with malaria and lymphatic filarial parasites, in archived and newly collected specimens from northeastern Tanzania," *Malaria Journal*, vol. 14, no. 1, p. 104, 2015.

[212] L. A. Kelly-Hope and F. E. McKenzie, "The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan Africa," *Malaria Journal*, vol. 8, no. 1, p. 19, 2009.

[213] M. A. Rider, B. D. Byrd, J. Keating, D. M. Wesson, and K. A. Caillouet, "PCR detection of malaria parasites in desiccated Anopheles mosquitoes is uninhibited by storage time and temperature," *Malaria Journal*, vol. 18, p. 314, 2012.

[214] E. Kamau, S. Alemayehu, K. C. Feghali, D. Saunders, and C. F. Ockenhouse, "Multiplex qPCR for Detection and Absolute Quantification of Malaria," *PLOS ONE*, vol. 8, p. e71539, 8 2013.

[215] P. M. Esperança, A. M. Blagborough, D. F. Da, F. E. Dowell, and T. S. Churcher, "Detection of Plasmodium berghei infected Anopheles stephensi using near-infrared spectroscopy," *Parasites and Vectors*, vol. 11, p. 377, 2018.

[216] P. Heraud, P. Chatchawal, M. Wongwattanakul, P. Tippayawat, C. Doerig, P. Jearanaikoon, D. Perez-Guaita, and B. R. Wood, "Infrared spectroscopy coupled to cloud-based data management as a tool to diagnose malaria: A pilot study in a malaria-endemic country," *Malaria Journal*, vol. 18, p. 348, 2019.

[217] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[218] H. S. Ngowo, E. W. Kaindoa, J. Matthiopoulos, H. M. Ferguson, and F. O. Okumu, "Variations in household microclimate affect outdoor-biting behaviour of malaria vectors," *Wellcome Open Research*, vol. 2, p. 102, 2017.

[219] M. F. Maia, M. G. Wagah, J. Karisa, R. Mwakesi, F. Mure, M. Muturi, J. Wambua, M. Hamaluba, F. E. Dowell, P. Bejon, and M. C. Kapulu, "Evaluation of near infrared spectroscopy for sporozoite detection in mosquitoes infected with wild-strain parasites from asymptomatic gametocyte carriers in Kilifi Kenya," *bioRxiv*, p. 2020.07.25.220830, 1 2020.

[220] K. J. Robson, U. Frevert, I. Reckmann, G. Cowan, J. Beier, I. G. Scragg, K. Takehara, D. H. Bishop, G. Pradel, and R. Sinden, "Thrombospondin-related adhesive protein (TRAP) of Plasmodium falciparum: expression during sporozoite ontogeny and binding to human hepatocytes.," *The EMBO Journal*, vol. 14, pp. 3883–3894, 8 1995.

[221] R. E. Sinden, "The cell biology of malaria infection of mosquito: Advances and opportunities," *Cellular Microbiology*, vol. 17, no. 4, pp. 451–466, 2015.

[222] A. Rivero and H. M. Ferguson, "The energetic budget of Anopheles stephensi infected with Plasmodium chabaudi: is energy depletion a mechanism for virulence?," *Proc R Soc Lond B Biol Sci*, vol. 270, no. 1522, pp. 1365–1371, 2003.

[223] Y. O. Zhao, S. Kurscheid, Y. Zhang, L. Liu, L. Zhang, K. Loeliger, and E. Fikrig, "Enhanced survival of plasmodium-infected mosquitoes during starvation," *PLoS ONE*, vol. 7, p. e40556, 2012.

[224] A. L. Hendershot, E. Esayas, A. C. Sutcliffe, S. R. Irish, E. Gadisa, F. G. Tadesse, and N. F. Lobo, "A comparison of PCR and ELISA methods to detect different stages of Plasmodium vivax in Anopheles arabiensis," *Parasites & Vectors*, vol. 14, no. 1, p. 473, 2021.

[225] A. T. DeGaetano, "Meteorological effects on adult mosquito (Culex) populations in metropolitan New Jersey," *International Journal of Biometeorology*, vol. 49, pp. 345–353, 2005.

[226] M. F. Maia, A. Robinson, A. John, J. Mgando, E. Simfukwe, and S. J. Moore, "Comparison of the CDC Backpack aspirator and the Prokopack aspirator for sampling indoor-and outdoor-resting mosquitoes in southern Tanzania," *Parasites and Vectors*, vol. 4, no. 1, pp. 1–10, 2011.

[227] L. E. Mboera, J. Kihonda, M. A. Braks, and B. G. Knols, "Short report: Influence of centers for disease control light trap position, relative to a human-baited bed net, on catches of Anopheles gambiae and Culex quinquefasciatus in Tanzania," *Am J Trop Med Hyg*, vol. 59, no. 4, pp. 595–596, 1998.

[228] K. S. Kreppel, P. C. D. Johnson, N. J. Govella, M. Pombi, D. Maliti, and H. M. Ferguson, "Comparative evaluation of the Sticky-Resting-Box-Trap, the standardised resting-bucket-trap and indoor aspiration for sampling malaria vectors," *Parasites & Vectors*, vol. 8, no. 1, p. 462, 2015.

[229] E. P. Mwanga, I. S. Mchola, F. E. Makala, I. H. Mshani, D. J. Siria, S. H. Mwinyi, S. Abbasi, G. Seleman, J. N. Mgaya, M. G. Jiménez, K. Wynne, M. T. Sikulu-Lord, P. Selvaraj, F. O. Okumu, F. Baldini, and S. A. Babayan, "Rapid assessment of the blood-feeding histories of wild-caught malaria mosquitoes using mid-infrared spectroscopy and machine learning," *Malaria Journal*, vol. 23, no. 1, p. 86, 2024.

[230] P. Chomczynski, K. Mackey, R. Drews, and W. Wilfinger, "DNAzol®: A Reagent for the Rapid Isolation of Genomic DNA," *BioTechniques*, vol. 22, pp. 550–553, 3 1997.

[231] T. Schindler, T. Robaina, J. Sax, J. R. Bieri, M. Mpina, L. Gondwe, L. Acuche, G. Garcia, C. Cortes, C. Maas, and C. Daubenberger, "Molecular monitoring of the diversity of human pathogenic malaria species in blood donations on Bioko Island, Equatorial Guinea," *Malaria Journal*, vol. 18, no. 1, p. 9, 2019.

[232] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, ICML United States, 2003.

[233] G. MacDonald, *The epidemiology and control of malaria*. London: Oxford University Press, 1957.

[234] M. Rougemont, M. Van Saanen, R. Sahli, H. P. Hinrikson, J. Bille, and K. Jaton, "Detection of four Plasmodium species in blood from humans by 18S rRNA gene subunit-based and species-specific real-time PCR assays," *Journal of Clinical Microbiology*, vol. 42, no. 12, pp. 5636–5643, 2004.

[235] C. Beadle, G. W. Long, P. D. McElroy, S. L. Hoffman, G. W. Long, W. R. Weiss, S. M. Maret, and A. J. Oloo, "Diagnosis of malaria by detection of Plasmodium falciparum HRP-2 antigen with a rapid dipstick antigen-capture assay," *The Lancet*, 1994.

[236] A. Moody, "Rapid diagnostic tests for malaria parasites," 2002.

[237] M. L. Wilson, "Malaria rapid diagnostic tests," *Clinical Infectious Diseases*, 2012.

[238] C. Drakeley and H. Reyburn, "Out with the old, in with the new: the utility of rapid diagnostic tests for malaria diagnosis in Africa," *Trans R Soc Hyg Trop Med*, vol. 103, no. 4, pp. 333–337, 2009.

[239] A. F. Fagbamigbe, "On the discriminatory and predictive accuracy of the RDT against the microscopy in the diagnosis of malaria among under-five children in Nigeria," *Malaria Journal*, vol. 18, no. 1, p. 46, 2019.

[240] L. Okell, A. Ghani, E. Lyons, and C. Drakeley, " Submicroscopic Infection in Plasmodium falciparum –Endemic Populations: A Systematic Review and Meta-Analysis ," *The Journal of Infectious Diseases*, 2009.

[241] P. Pati, G. Dhangadamajhi, M. Bal, and M. Ranjit, "High proportions of pfhrp2 gene deletion and performance of HRP2-based rapid diagnostic test in Plasmodium falciparum field isolates of Odisha," *Malaria Journal*, vol. 17, no. 1, p. 394, 2018.

[242] 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, no. 1, p. 422, 2020.

[243] 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," *bioRxiv*, p. 490326, 1 2018.

[244] E. P. Mwanga, D. J. Siria, I. H. Mshani, S. H. Mwinyi, S. Abbasi, M. G. Jimenez, K. Wynne, F. Baldini, S. A. Babayan, and F. O. Okumu, "Rapid classification of epidemiologically relevant age categories of the malaria vector, Anopheles funestus," *Parasites & Vectors*, vol. 17, no. 1, p. 143, 2024.

[245] 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, 4 2021.

[246] 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, p. eaat0496, 5 2018.

[247] 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, pp. 5285–5293, 5 2017.

[248] L. M. B. 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, no. 1, p. 67, 2021.

[249] E. P. Mwanga, P. A. Kweyamba, D. J. Siria, I. H. Mshani, I. S. Mchola, F. E. Makala, G. Seleman, S. Abbasi, S. H. Mwinyi, M. González-Jiménez, K. Waynne, F. Baldini, S. A. Babayan, and F. O. Okumu, "Reagent-free detection of Plasmodium falciparum malaria infections in field-collected mosquitoes using mid-infrared spectroscopy and machine learning," *Scientific Reports*, vol. 14, no. 1, p. 12100, 2024.

[250] F. E. Dowell, A. E. M. Noutcha, and K. Michel, "The Effect of Preservation Methods on Predicting Mosquito Age by Near Infrared Spectroscopy," *The American Society of Tropical Medicine and Hygiene*, vol. 85, no. 6, pp. 1093–1096, 2011.

[251] V. S. Mayagaya, A. J. Ntamatungiro, S. J. Moore, R. A. Wirtz, F. E. Dowell, and M. F. 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.

[252] 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 s.l. age and species," *Malar J*, vol. 10, p. 186, 2011.

[253] R. J. Golden, S. E. Holm, D. E. Robinson, P. H. Julkunen, and E. A. Reese, "Chloroform Mode of Action: Implications for Cancer Risk Assessment," *Regulatory Toxicology and Pharmacology*, vol. 26, no. 2, pp. 142–155, 1997.

[254] B. M. Somé, D. F. Da, R. McCabe, N. D. C. Djègbè, L. I. G. Paré, K. Wermé, K. Mouline, T. Lefèvre, A. G. Ouédraogo, T. S. Churcher, and R. K. Dabiré, "Adapting field-mosquito collection techniques in a perspective of near-infrared spectroscopy implementation," *Parasites & Vectors*, vol. 15, no. 1, p. 338, 2022.

[255] M. Jackson, K. Kim, J. Tetteh, J. R. Mansfield, B. Dolenko, R. L. Somorjai, F. W. Orr, P. H. Watson, and H. H. Mantsch, "Cancer diagnosis by infrared spectroscopy: methodological aspects," in *Proc.SPIE*, vol. 3257, pp. 24–34, 4 1998.

[256] J. N. Mgaya, D. J. Siria, F. E. Makala, J. P. Mgando, J.-M. Vianney, E. P. Mwanga, and F. O. Okumu, "Effects of sample preservation methods and duration of storage on the performance of mid-infrared spectroscopy for predicting the age of malaria vectors," *Parasites & Vectors*, vol. 15, no. 1, p. 281, 2022.

[257] R. P. Penilla, M. H. Rodriguez, A. D. Lopez, J. M. Viader-Salvado, and C. N. Sanchez, "Pteridine concentrations differ between insectary-reared and field-collected Anopheles albimanus mosquitoes of the same physiological age," *Medical and Veterinary Entomology*, vol. 16, pp. 225–234, 9 2002.

[258] 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, pp. 811–823, 11 1999.

[259] L. E. Hugo, G. K. Eaglesham, B. H. Kay, P. A. Ryan, and N. Holling, "Investigation of cuticular hydrocarbons for determining the age and survivorship of Australasian mosquitoes," *The American Journal of Tropical Medicine and Hygiene*, vol. 74, pp. 462–474, 3 2006.

[260] 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, pp. 900–912, 8 2020.

[261] 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, p. 20191091, 7 2019.

[262] Bruker Optics, "ALPHA II - The Compact FTIR Spectrometer for any Industry," 2019.

[263] Bruker Optics, "OPUS Spectroscopy Software," 2019.

[264] M. González-Jiménez, "A custom program that imports the IR spectra, cleans and screens them to eliminate the badly measured ones, and extracts the most interesting data from them!," 2019.

[265] I. H. Mshani, F. M. Jackson, R. Y. Mwanga, P. A. Kweyamba, E. P. Mwanga, M. M. Tambwe, L. M. Hofer, D. J. Siria, M. González-Jiménez, K. Wynne, S. J. Moore, F. Okumu, S. A. Babayan, and F. Baldini, "Screening of malaria infections in human blood samples with varying parasite densities and anaemic conditions using AI-Powered mid-infrared spectroscopy," *Malaria Journal*, vol. 23, no. 1, p. 188, 2024.

[266] B. J. Johnson, L. E. Hugo, T. S. Churcher, O. T. W. Ong, and G. J. Devine, "Mosquito Age Grading and Vector-Control Programmes," *Trends in Parasitology*, vol. 36, no. 1, pp. 39–51, 2020.

[267] S. Niare, J. M. Berenger, C. Dieme, O. Doumbo, D. Raoult, P. Parola, and L. Almeras, "Identification of blood meal sources in the main African malaria mosquito vector by MALDI-TOF MS," *Malaria Journal*, 2016.

[268] S. Niare, L. Almeras, F. Tandina, A. Yssouf, A. Bacar, A. Toilibou, O. Doumbo, D. Raoult, and P. Parola, "MALDI-TOF MS identification of Anopheles gambiae Giles blood meal crushed on Whatman filter papers," *PLoS ONE*, 2017.

[269] F. Tandina, M. Laroche, B. Davoust, O. K Doumbo, and P. Parola, "Blood meal identification in the cryptic species Anopheles gambiae and Anopheles coluzzii using MALDI-TOF MS," *Parasite*, 2018.

[270] D. E. Goldberg, A. F. Slater, A. Cerami, and G. B. Henderson, "Hemoglobin degradation in the malaria parasite Plasmodium falciparum: an ordered process in a unique organelle.," *Proceedings of the National Academy of Sciences*, vol. 87, pp. 2931–2935, 4 1990.

[271] F. Okumu and M. Finda, "Key Characteristics of Residual Malaria Transmission in Two Districts in South-Eastern Tanzania—Implications for Improved Control," *The Journal of Infectious Diseases*, vol. 223, pp. S143–S154, 4 2021.

[272] B. J. Msugupakulya, S. K. Ngajuma, A. N. Ngayambwa, B. E. Kidwanga, I. R. Mpasuka, P. Selvaraj, A. L. Wilson, and F. O. Okumu, "Influence of larval growth and habitat shading on retreatment frequencies of biolarvicides against malaria vectors," *Scientific Reports*, vol. 14, p. 1002, 1 2024.

[273] V. S. Mayagaya, G. Nkwengulila, I. N. Lyimo, J. Kihonda, H. Mtambala, H. Ngonyani, T. L. Russell, and H. M. Ferguson, "The impact of livestock on the abundance, resting behaviour and sporozoite rate of malaria vectors in southern Tanzania," *Malaria Journal*, vol. 14, no. 1, 2015.