Near Infrared Spectroscopy for Estimating the Age of Malaria Transmitting Mosquitoes

Masabho Peter Milali
Marquette University

Recommended Citation
http://epublications.marquette.edu/theses_open/377
NEAR INFRARED SPECTROSCOPY FOR ESTIMATING THE AGE OF MALARIA TRANSMITTING MOSQUITOES

by

Masabho Peter Milali, B.S.

A Thesis Submitted to the Faculty of the Graduate School, Marquette University, in Partial Fulfillment of the Requirements for the Degree of Master of Science

Milwaukee, Wisconsin

December 2016
ABSTRACT

NEAR INFRARED SPECTROSCOPY FOR ESTIMATING THE AGE OF MALARIA TRANSMITTING MOSQUITOES

Masabho Peter Milali, B.S.
Marquette University, 2016

We explore the use of near infrared spectrometry to classifying the age of a wild malaria transmitting mosquito. In Chapter Two, using a different set of lab-reared mosquitoes, we replicate the Mayagaya et al. study of the accuracy of near-infrared spectrometry (NIRS) to estimate the age of lab-reared mosquitoes, reproducing the published accuracy. Our results strengthen the Mayagaya et. al study and increase confidence in using NIRS to estimate age classes of mosquitoes.

In the field, we wish to classify the ages of wild, not lab-reared mosquitoes, but the necessary training data from wild mosquitoes is difficult to find. Applying a model trained on spectra from lab-reared mosquitoes to estimate the age of wild mosquitoes is appropriate only if spectra collected from lab-reared mosquitoes are equivalent to those collected from wild mosquitoes. In Chapter Three, we apply k means cluster analysis to a mixture of spectra collected from lab-reared and wild Anopheles arabiensis mosquitoes to determine if there is significant difference between these spectra. We find no significant difference ($P = 0.245$) in distributions between the wild and lab-reared mosquitoes in the two formed clusters. The two formed clusters have average silhouette coefficient values (cluster quality measure) of 0.51 and 0.77, respectively, which shows that the clusters were reasonable and strong, respectively.

Basing on results from Chapter Three, we estimate the age class of wild Anopheles arabiensis mosquitoes using a classification model trained on lab-reared Anopheles arabiensis. We validate the accuracy of the model by comparing its estimates with ovary dissection estimates. While our model estimated 86% and 14% of wild Anopheles arabiensis to be $< 7$ and $\geq 7$ days old, respectively, ovary dissection estimated 72% as young and 28% as old. Studies show that wild mosquito populations generally consist of more young than old mosquitoes. Therefore, our model estimates age of wild mosquitoes in consistency with ovary dissection and other studies conducted to determine age structure of wild mosquitoes.
ACKNOWLEDGMENTS

Masabho Peter Milali, B.S.

I would like to express my sincere appreciation to my principal supervisor, Dr. George Corliss, for his constant guidance and encouragement, without which this work would not have been possible. For his unwavering support, I am truly grateful. I am also grateful to my committee members, Dr. Richard Povinelli, Dr. Samson Kiware and Dr. Stephen Merrill, for their support towards the successful completion of this thesis.

Without the financial support from the Marquette Graduate School, which offered me a scholarship for graduate studies, this work would not have been possible. Special thanks go to Marquette MSCS department for accepting my application, which lead me granted this rare opportunity. I also would like to express my heartfelt gratitude to the Ifakara Health Institute, specifically Dr. Dickson Lwetojeira, Dr. Caroline Harris, and Dr. Silas Majambere, who initially employed me, an employment which lead me to meet Dr. George Corliss and Dr. Samson Kiware, who closely supported and facilitated the whole application process to join Marquette University. Many thanks also go to Dr. Maggy Sikulu and Dr. Fredros Okumu, who later accepted me into their project, which not only lead me into my master thesis topic, but also allowed me to use the data. I am grateful to them.

I would also like to thank the GasDay lab, Dr. Ronald Brown, Dr. George Corliss, Dr. Richard Povinelli, Thomas Quinn, and Catherine Porter for putting GasDay resources at my disposal during my master study. To my colleagues and friends in the lab, thank you so much for the moral support, which made my studies more enjoyable, despite endless life and school challenges and stresses.

Finally, to my parents, especially my mother, Daines Milali, who has always believed in me and worked tirelessly to make sure I become a man of value and successful, thank you Mom. To my family Georgia and Daniella, we did it! Thank you for your understanding and support you have given me; I am glad to have you.
# TABLE OF CONTENTS

**ACKNOWLEDGMENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

**CHAPTER 1  Contribution of Mosquito Age Classification toward Eliminating Malaria Transmission**

1.1 Importance in the Fight Against Malaria of Knowing Mosquito Age . 1
1.2 Statement of the Problem . . . . . . . . . . . . . . . . . . . . . . . . 2
1.3 Current Status of the Problem . . . . . . . . . . . . . . . . . . . . . 3
1.4 Thesis Specific Objectives . . . . . . . . . . . . . . . . . . . . . . . 4
1.5 Statement of the Approach Used . . . . . . . . . . . . . . . . . . . . 5

**CHAPTER 2  Reproducibility of the Accuracy of Near Infrared Spectroscopy for Estimating the Age of Lab-Reared Mosquitoes**

2.1 Mosquito Age and Malaria Transmission 9
2.2 Mosquitoes and Spectra Collection . . . . . . . . . . . . . . . . . . 11
2.3 Mayagaya et al. Model Training . . . . . . . . . . . . . . . . . . . . 14
2.4 Training Our Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
    2.4.1 The Actual Age Prediction Model . . . . . . . . . . . . . . . . . 18
    2.4.2 Classification Models A and B . . . . . . . . . . . . . . . . . . 21
2.5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

**CHAPTER 3  Do NIR Spectra Collected from Lab-reared Mosquitoes Differ from Those Collected from Wild Mosquitoes?**

3.1 Mosquito Age and Malaria Transmission . . . . . . . . . . . . . . . . 29
3.2 Overview of Cluster Analysis . . . . . . . . . . . . . . . . . . . . . . 32
3.3 Mosquitoes and Spectra Collection .................................. 34
3.4 \(K\)-means Clustering on a Mixed Spectra ............................. 36
3.5 Discussion - What do results from this study mean? ................. 43
3.6 Conclusion - No Clear Difference Between Spectra Collected from Lab-reared and Wild Mosquitoes of the Same Species ............. 49

CHAPTER 4 Age Classification of Wild Mosquitoes Using a Classification Model Trained on Lab-reared Mosquitoes (Model B in Chapter Two) 52

4.1 Age-class Estimation of Wild Anopheles arabiensis Using Model B Developed from Lab-reared Anopheles arabiensis in Chapter Two .... 52
4.2 Spectra Collection and Ovary Dissection of Wild An. arabiensis .... 55
4.3 Age Classification of Wild Anopheles arabiensis ....................... 57
4.4 Analysis on the Number of Nulliparous and Parous Mosquitoes in Each Age-class Obtained from Model B ................................. 59
4.5 Comparison Between Number of Mosquitoes in Each Age Class Obtained when Model B and Detinova Ovary Dissection Are Separately Used to Classify Age of the Same Set of Mosquitoes .................. 62
4.6 Comparison Between Model B Age Classification and Other Studies on Age Structure of Wild Mosquitoes ............................. 65
4.7 Model B Trained on Lab-reared Mosquitoes Correctly Classifies the Age of Wild Mosquitoes ........................................ 68

CHAPTER 5 No Clear Difference on Spectra Collected from Lab-reared and Wild Mosquitoes: A Classification Model Trained on Spectra from Lab-reared Mosquitoes Classifies Wild Mosquitoes Consistently with Classification by Detinova Ovary Dissection 69

5.1 Summary of Thesis Findings and Contributions to the Fight Against Malaria ............................................................ 69
5.2 Recommendations for Future Studies .................................... 71

Bibliography

BIBLIOGRAPHY
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Mayagaya et al. percent accuracy predicting the age of <em>Anopheles gambiae</em> mosquitoes into specific age groups when using a calibration developed from the Ifakara strain of <em>An. gambiae</em> mosquitoes [14].</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Mayagaya et al. accuracy of mosquito age prediction (in days) when using a partial least squares regression cross-validation model (within-strain prediction) of Ifakara <em>Anopheles gambiae</em> s.s strain mosquitoes [14].</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>Our percentage accuracy of predicting the age of <em>Anopheles gambiae</em> mosquitoes into specific age groups when using a regression model developed from the Ifakara strain of <em>An. gambiae</em> mosquitoes.</td>
<td>24</td>
</tr>
<tr>
<td>2.4</td>
<td>Our average actual age prediction (in days) when using a partial least squares regression cross-validation model (within-strain prediction) of Ifakara <em>Anopheles gambiae</em> s.s strain mosquitoes.</td>
<td>24</td>
</tr>
<tr>
<td>2.5</td>
<td>Our accuracy of predicting the age of <em>Anopheles gambiae</em> mosquitoes into specific age groups when using classification models developed from the Ifakara strain of <em>An. gambiae</em> mosquitoes.</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Interpretation of the silhouette values for partitioning methods [29].</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Distribution of mosquitoes in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c.X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$), and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise) when all ages of lab-reared mosquitoes were included in the analysis.</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Distribution of spectra in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c.X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$) and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise), when the ages of lab-reared were controlled before clustering.</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Distribution of spectra in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c.X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$) and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise), when 3, 5, and 25 day old lab-reared mosquitoes were not included in the analysis.</td>
<td>51</td>
</tr>
</tbody>
</table>
4.1 Age classification of wild *Anopheles arabiensis* by model B. . . . . . . 58

4.2 The number of mosquitoes in each age class of wild mosquitoes obtained when model B and Detinova ovary dissection are independently applied to classify the age of the same population of mosquitoes. . . . 64
LIST OF FIGURES

1.1 Age-class estimations of wild *Anopheles arabiensis* mosquitoes by a NIRS classification model trained on lab-reared *Anopheles arabiensis* compared to the age-class estimations by Detinova ovary dissection [7]. 8

2.1 Mosquitoes are scanned using near-infrared spectrometer. A) Plate with killed mosquitoes positioned for scanning. B) Spectra from the mosquitoes. C) Complete NIRS system (ASD Inc., Boulder, CO) [14]. 13

2.2 Shows our spectra before and after cleaning. A) Spectra before cleaning (n = 786), B) Spectra after cleaning noisy ends (n = 786), C) Spectra after discarding spectra with absorbance range < 0.3 (n = 785), D) Spectra after discarding spectra with absorbances either < 0.4, > 1.4, or both (n = 442). 17

2.3 Regression coefficients used by our regression model to estimate ages of mosquitoes. This figure corresponds to Figure 6 in Mayagaya et al. [14]. 21

2.4 Estimated ages from our regression model vs. actual mosquito ages on test data. This figure corresponds to Figure 5 in Mayagaya et al. [14]. 22

2.5 Box plot showing estimated ages from our regression model vs. actual mosquito ages on test data. 23

2.6 Regression coefficients used by classification model A to classify ages of mosquitoes. 25

2.7 Regression coefficients used by classification model B to classify ages of mosquitoes. 26

3.1 Two-dimensional plot of clusters using absorbances at 350 nm and 351 nm, when all ages of lab-reared mosquitoes were used in the analysis. 37

3.2 Silhouette coefficient of each spectrum (object) in its associated cluster when we mixed all ages of lab-reared with wild *Anopheles Arabiensis*. The average SC for cluster one and two is 0.77 and 0.53 indicating the clusters are strong and reasonable, respectively. 38

3.3 Silhouette coefficients of each spectrum (object) in its associated cluster when we mixed all ages of lab-reared with wild *Anopheles Arabiensis*. 39
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Two dimensional plot of clusters using absorbances at 350nm and 351nm when 3, 5, and 25 day old lab-reared mosquitoes were removed during cluster analysis.</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Silhouette coefficient of each spectrum (object) in its associated cluster when 3, 5, and 25 day old lab-reared mosquitoes were not included in the analysis.</td>
<td>43</td>
</tr>
<tr>
<td>3.6</td>
<td>Silhouette coefficient of each spectrum (object) in its associated cluster when 3, 5, and 25 day old lab-reared mosquitoes were removed during cluster analysis.</td>
<td>44</td>
</tr>
<tr>
<td>3.7</td>
<td>Two dimensional plot of clusters using first and second partial least squares (PLS) components.</td>
<td>45</td>
</tr>
<tr>
<td>3.8</td>
<td>Silhouette coefficients of each spectrum (object) in its associated cluster when cluster analysis uses ten PLS components.</td>
<td>46</td>
</tr>
<tr>
<td>3.9</td>
<td>Silhouette coefficients of each spectrum (object) in its associated cluster when cluster analysis uses ten PLS components.</td>
<td>47</td>
</tr>
<tr>
<td>4.1</td>
<td>Process of wild mosquito collection, spectra collection, ovary dissection, and species identification.</td>
<td>54</td>
</tr>
<tr>
<td>4.2</td>
<td>Spectra collection of wild <em>Anopheles gambiae</em> complex mosquitoes before ovary dissection and species identification by polymerase chain reaction.</td>
<td>56</td>
</tr>
<tr>
<td>4.3</td>
<td>Ovary dissection of wild <em>Anopheles gambiae</em> complex mosquitoes before their species identification by PCR. A) Dissected mosquitoes; B) Mosquito Dissection process; C) Zoomed in dissection process.</td>
<td>57</td>
</tr>
<tr>
<td>4.4</td>
<td>Percentage distribution of nulliparous and parous mosquitoes in each age-class from model B.</td>
<td>60</td>
</tr>
<tr>
<td>4.5</td>
<td>Percentage of mosquitoes in each age class of wild <em>Anopheles arabiensis</em> obtained when model B and Detinova ovary dissection are separately applied to classify the age of the same population of mosquitoes (N = 927).</td>
<td>63</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison on the percentage of mosquitoes in each age class obtained when model B and Detinova ovary dissection are independently applied to classify the age of the same set of wild <em>Anopheles arabiensis</em> per month of collection.</td>
<td>65</td>
</tr>
</tbody>
</table>
4.7 Modified from Brownstein et al. [5] showing age structure of a simulated wild mosquito population. Yellow bars represent mosquitoes which are less likely to be infectious, and red bars represent potential infectious mosquitoes.
CHAPTER 1

Contribution of Mosquito Age Classification toward Eliminating Malaria Transmission

This chapter gives a general overview of the thesis. It provides a general statement of the problem, the contributions of this thesis toward finding solutions to the problem, methods used, main results obtained, and conclusions drawn from the study.

1.1 Importance in the Fight Against Malaria of Knowing Mosquito Age

Estimating the age of mosquitoes is one of the methods entomologists use as a leading indicator of the rate of malaria transmission in an area and the effectiveness of an existing mosquito population control method. Malaria is a vector-borne parasitic disease transmitted to people by mosquitoes of the genus Anopheles. The disease killed approximately 438,000 people in 2015 [35]. Mosquitoes contribute to malaria transmission by hosting and allowing the development to maturity of the malaria-causing Plasmodium parasite [3]. Female mosquitoes acquire Plasmodium when they feed on infected human blood for their egg development. Depending on environmental temperature, Plasmodium takes 10-14 days in an Anopheles mosquito to develop fully enough to cause malaria in humans [3]. This means that if a
mosquito is less than 10 days old, the chances that it is carrying a full developed parasite (which is considered infectious) are small. Therefore, knowing the age of a mosquito helps determine whether a mosquito is capable of transmitting malaria.

Knowing the age of a mosquito also is important when evaluating an intervention to control a population of mosquitoes. Vector control interventions such as the use of insecticide treated nets (ITNs) and indoor residual spraying (IRS) normally are implemented in areas known to have a high population of malaria vectors (mosquitoes). Once implemented, the interventions are expected to reduce mosquito abundance and lifespan to a level that does not support *Plasmodium* parasite development to maturity. Monitoring and evaluation of these vector control intervention (ITNs and IRS) involves determining age and species composition of the mosquito population before and after intervention. The presence of a small number of old mosquitoes in an area with an intervention (ITNs or IRS) indicates that the method (intervention) is working. On the other hand, if there are more old mosquitoes, it shows that the intervention is not working.

1.2 Statement of the Problem

The current techniques used to estimate mosquito age are based on dissecting her ovaries by hand to determine whether she has laid eggs. Those found to have laid eggs are assumed to be older than those found to not have laid eggs [7]. This
assumption can be misleading, as mosquitoes can be old but have not laid eggs and can be three to six days old, which is considered young, and have laid eggs. The method is laborious, difficult, and limited to only few experts. As a result, we need a new approach to address these limitations.

Near Infrared Spectrometry (NIRS) can be a complementary method to mosquito ovary dissection. NIRS is a high throughput, automated technique, which measures the amount of the near infrared energy absorbed by samples. NIRS has been applied to identify species of insects infecting stored grains [9]; to age grade houseflies [20], stored-grain pests [21], and biting midges [22]; to differentiate between species and subspecies of termites [1]; to estimate the age and to identify species of morphologically indistinguishable laboratory reared and semi-field raised *Anopheles gambiae* and *Anopheles arabiensis* mosquitoes [14, 26]; and to detect and identify two strains of *Wolbachia pipientis* (wMelPop and wMel) in male and female laboratory-reared *Aedes aegypti* mosquitoes [28]. The study by Mayagaya et al. [14] reports that NIRS can classify the age of lab-reared and semi-field mosquitoes into either less than 7 or greater than 7 days old with an accuracy exceeding 80%.

1.3 Current Status of the Problem

The ability of NIRS to estimate the age of laboratory and semi-field raised mosquitoes is not what is needed. Lab-reared mosquitoes do not transmit malaria;
wild mosquitoes do. It is unknown if NIRS can estimate the age of wild mosquitoes accurately. The limitation is the lack of age-labeled wild mosquitoes with which to train a model. Training a model using labels from ovary dissection yields a model with poor accuracy. Applying a model trained on spectra from lab-reared mosquitoes to estimate the age of wild mosquitoes would be an appropriate only if spectra collected from lab-reared mosquitoes are equivalent to those collected from wild mosquitoes, but no studies have validated that generalization. This thesis explores this generalization by performing $k$-means cluster analysis on a mixture of spectra collected from lab-reared and wild *Anopheles arabiensis* to determine if there is any significant difference.

### 1.4 Thesis Specific Objectives

The thesis has three specific objectives:

1. To replicate the study of Mayagaya et al. [14] on the accuracy of NIRS to estimate the age of lab-reared mosquitoes.

2. To determine if there are spectral differences between lab-reared and wild mosquitoes of the same species

3. To estimate the age of wild *Anopheles arabiensis* mosquitoes using a classification model trained on lab-reared *Anopheles arabiensis* and compare
the model estimations with the parity status of the wild *Anopheles arabiensis*

obtained from ovary dissection.

The thesis is organized in such a way that each specific objective is an independent chapter with a stand-alone introduction, methods, results, and conclusions.

Objectives one and two are discussed in Chapters Two and Three, respectively, and are in preparation for submission to a peer-reviewed journal. Objective three is considered in Chapter Four of the thesis. Chapter Five presents general conclusions and suggestions for future research work.

1.5 Statement of the Approach Used

Is the published ability and accuracy of near infrared spectroscopy to estimate the age of lab-reared mosquitoes reproducible? In Chapter Two, we test the reproducibility of the accuracy of near infrared spectrometry for estimating the age of lab-reared mosquitoes as published by Mayagaya et al. in 2009 [14]. We replicate the study using a different set of spectra collected from lab-reared *Anopheles gambiae* mosquitoes. While Mayagaya et al. trained a regression model and interpreted results as classification, we extend their study by training proper classification models and compare the results. We find that the accuracy of near infrared spectrometry for estimating the age of lab-reared *Anopheles gambiae* as
published by Mayagaya et al. 2009 is reproducible. When the proper classification model was trained to classify lab-reared mosquito age as young ($< 7$) or old ($\geq 7$ days), the accuracy using our data was 83.5%, compared to the Mayagaya et al. model accuracy of $\geq 80\%$.

Is there a significant difference in spectra collected from lab-reared and wild mosquitoes of the same species? Following the successful replication of the ability of near infrared spectrometry to classify ages of lab-reared mosquitoes as published by Mayagaya et al. 2009, in Chapter Three, we perform $k$-means cluster analysis on a mixture of spectra collected from lab-reared and wild *Anopheles arabiensis* to determine if there is any significant spectral difference. We hypothesized that from a spectra point of view, there is no significant difference between spectra collected from lab-reared mosquitoes and those from wild mosquitoes. We find two clusters with average silhouette coefficients of 0.51 and 0.77, respectively. The silhouette coefficient, explained in Chapter Three, helps evaluate the quality of the formed clusters before we perform any statistical analysis. A $X^2$ analysis showed no significant difference in distribution of spectra collected from lab-reared and wild mosquitoes between the formed clusters ($P = 0.245$). We suspect clustering was due to age differences of mosquitoes and not their source. These results strengthened the idea of training the model to estimate the age of wild mosquitoes using spectra collected from lab-reared mosquitoes.
Can the classification model trained on lab-reared mosquitoes be applied to estimate the age-class of wild mosquitoes of the same species? After our cluster analysis to determine similarity between spectra collected from lab-reared and wild mosquitoes of the same species had failed, in Chapter Four we estimate the age class (either $\leq 7$ or $\geq 7$ days) of wild *Anopheles arabiensis* mosquitoes using the classification model trained (in Chapter Two) on lab-reared *Anopheles arabiensis* and compared the age-class estimates from the model with the age-class estimates from Detinova ovary dissection. As summarized in Figure 1.1, the age class estimations of wild *Anopheles arabiensis* by a classification model trained on lab-reared *Anopheles arabiensis* was consistent with age-class estimations by ovary dissection (Detinova).

In summary, we find that there is high chance that a model trained using spectra collected from lab-reared mosquitoes can be applied to estimate the age class of wild mosquitoes. Further studies are recommended to validate the accuracy of the classification model trained on lab-reared and applied to estimate the age class of wild mosquitoes.
Figure 1.1: Age-class estimations of wild *Anopheles arabiensis* mosquitoes by a NIRS classification model trained on lab-reared *Anopheles arabiensis* compared to the age-class estimations by Detinova ovary dissection [7].
CHAPTER 2

Reproducibility of the Accuracy of Near Infrared Spectroscopy for Estimating the Age of Lab-Reared Mosquitoes

This chapter presents results obtained when a different mosquito data set was used to determine how reproducible is the published [14] accuracy of near-infrared spectrometry to estimate the age of lab-reared mosquitoes. This chapter is a preliminary version of [16] in preparation for submission to a peer-reviewed journal.

2.1 Mosquito Age and Malaria Transmission

Knowing the age of a mosquito is very important in determining whether a mosquito is capable of transmitting malaria. Malaria parasites need at least 10 days to develop in a mosquito before they can be transmitted to humans [3]. Mosquitoes fewer than 10 days old cannot carry fully developed parasites, which can be transmitted to humans and cause malaria. To prioritize vector management, especially when resources are scarce, it is recommended to determine mosquito age structure before any intervention. An area with old mosquitoes is considered to have higher active malaria transmission than the area with young mosquitoes and needs quick intervention. Knowing the age of a mosquito also helps to monitor and evaluate the existing intervention in a particular area. The presence of a large
number of old mosquitoes in an area indicates that the existing method is not working. On the other hand, if there are more young mosquitoes, the existing controls are working.

The current techniques used to estimate mosquito age are based on hand dissection of their ovaries to determine whether they have laid eggs. Those found to have laid eggs are assumed to be older (on average) than those found not to have laid eggs [7]. However, the difficulty and laborious dissections involved with this technique limits its application to only a few experts working with small numbers of mosquitoes. In addition, this method cannot estimate the actual age; it is limited to age classification (age grouping). As a result, a new approach which can address these limitations is needed.

Near Infrared Spectrometry (NIRS) has been considered as an alternative to hand dissection. It is a high throughput, automated technique, which measures the amount of the near infrared energy absorbed by samples. NIRS initially was developed for identification of cryptic (hidden) species of insects found in stored grains [9]. It also has been tested for age grading houseflies [20], stored-grain pests [21], biting midges [22], and to differentiate between species and subspecies of termites [1]. For mosquitoes, it has been used to estimate the age and to identify the species of morphologically indistinguishable *Anopheles gambiae* and *Anopheles arabiensis* in the laboratory [14] and in semi-field environments [26].
Mayagaya et al. 2009 published results showing the effectiveness of near infrared spectrometry for estimating the age of lab-reared *Anopheles gambiae* mosquitoes. They report that NIRS can classify the age of female lab-reared *Anopheles gambiae* as young (< 7 days) or old (≥ 7 days) with an accuracy of ≥ 80%. Our study reproduced their results on a different data set. While Mayagaya et al. trained the regression model and interpreted results as classification, we reproduced their study and went further by training a proper classification model and compared the results. We found consistency in results between the Mayagaya et al. and our studies and between the two models (regression and classification models). This consistency in results strengthen the assertion of Mayagaya et al. for the value of NIRS for classifying the age of mosquitoes.

### 2.2 Mosquitoes and Spectra Collection

The data used in both studies (Mayagaya et al. and ours) were in spectra obtained after scanning lab-reared (raised) mosquitoes. Mayagaya et al. used spectra collected from *Anopheles gambiae* with ages of 1, 4, 7, 10, 13, 16, and 19 days. These mosquitoes were obtained from three different insectaries: Kansas State University (KSU), Manhattan, Kansas; Centers for Disease Control and Prevention (CDC), Atlanta, Georgia; and Ifakara Health Institute (IHI), Ifakara, Tanzania. Details on how these mosquitoes were raised (reared) can be found in [14]. Before
scanning (collection of spectra), they killed mosquitoes using chloroform and placed 20 dead mosquitoes on the spectralon plate. They scanned individual mosquitoes by rotating the plate until the head and thorax of the mosquito were positioned 2 mm below a 3 mm-diameter fiber optic NIR probe. They scanned using a QualitySpec Pro Spectrometer (350 - 2500 nm; ASD Inc, Boulder, CO, Figure 2.1) with ASD software RS3 version 3.1 installed. More information about the NIRS machine and scanning can be found in [22]. They scanned approximately 365 mosquitoes from each of the three insectaries and recorded the status of each mosquito (blood fed, unfed, or gravid (carrying eggs/pregnant)).

Our study used the same *Anopheles gambiae* mosquito species with ages 1, 3, 5, 7, 9, 11, 15, and 20 days from the Ifakara Health Institute insectary. This insectary started in 2010 and is reared in a semi-field system (SFS) [17] under ambient temperature and light-dark cycles. The humidity is artificially increased to 80% during the dry season (May - October). Adult mosquitoes are daily provided a human arm as a blood meal source and 10% glucose solution. The insectary keeps records of mosquitoes from egg laying to adult emergence, and the cages are labeled in such a way mosquito ages are easily identified. Using the same NIRS machine, we collected spectra exactly as in Mayagaya et al.; the only difference was how we killed our mosquitoes. While they killed their mosquitoes using chloroform, we killed ours by freezing for 20 minutes. We did not record status of mosquitoes (i.e.,
blood fed, unfed, or gravid). A total of 786 mosquitoes were scanned with at least
70 mosquitoes from each age group (1, 3, 5, 7, 9, 11, 15, and 20 days).

Figure 2.1: Mosquitoes are scanned using near-infrared spectrometer. A) Plate with
killed mosquitoes positioned for scanning. B) Spectra from the mosquitoes. C) Com-
plete NIRS system (ASD Inc., Boulder, CO) [14].
2.3 Mayagaya et al. Model Training

After collecting spectra from mosquitoes, Mayagaya et al. discarded all spectra with absorbance ranges less than 0.3 and “generally outside the 0.5-1.0 absorbance range.” Using Grams software PLSPlus/IQ, they performed partial least square regression (PLS) on these spectra for both feature reduction and model training. They used leave-one-out cross-validation to train the model [14]. They chose the number of features by examining the reductions in the residual sum of squares obtained when additional features were added to the cross-validation models and by calculating the accuracy of the model when predicting the test set. In the end, they used six features to train the regression model.

They analyzed the model output (results) in three different ways: first, on how it estimated the actual age of a mosquito; second, on how it classified mosquitoes into three age groups (≤ 7d, 10d, and ≥ 13d); and third, on how it classified mosquitoes into two age groups (< 7d and ≥ 7d).

When interpreting the output of the model estimating actual ages, mosquitoes with model forecast ages < 2.5 days were considered as 1 day old, 2.5-5.4 as 4 days old, 5.5-8.4 as 7 days old, 8.5-11.4 as 10 days old, 11.5-14.4 as 13 days old, 14.5-17.4 as 16 days old, and 17.5-20.5 as 19 days old (rule for continuous estimation to discrete classification). Table 2.1 summarizes the performance of the
Mayagaya et al. model. The model did not perform well. The highest percent accuracy (the number of mosquitoes correctly estimated over the total number of mosquitoes of that particular age) was 54% (when estimating 10 day old mosquitoes). For the remaining ages, the model accuracy did not go beyond 27%.

The model accuracy improved when classifying mosquitoes into age classes. When classifying into three age classes, only 7, 13, and 19 day old mosquitoes were poorly estimated with an accuracy less than 50%. When classifying mosquitoes into two age groups, the model accuracy became even better, with only 7 day old mosquitoes poorly estimated (14%). The remaining mosquito ages were estimated with an accuracy of at least 76%. Table 2.2 shows the means and standard deviations of the model age estimates. The results corroborates the results presented in Table 2.1. Only 10 day old mosquitoes had an estimation mean close to the actual age and a small standard deviation. The remaining ages had a mean of the estimation sharply different from the actual age.

Table 2.1: Mayagaya et al. percent accuracy predicting the age of *Anopheles gambiae* mosquitoes into specific age groups when using a calibration developed from the Ifakara strain of *An. gambiae* mosquitoes [14].

<table>
<thead>
<tr>
<th>Actual age (in days)</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>16</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual age</td>
<td>24%</td>
<td>14%</td>
<td>12%</td>
<td>54%</td>
<td>27%</td>
<td>14%</td>
<td>2%</td>
</tr>
<tr>
<td>≤ 7,10, ≥ 13</td>
<td>90%</td>
<td>76%</td>
<td>14%</td>
<td>54%</td>
<td>27%</td>
<td>73%</td>
<td>44%</td>
</tr>
<tr>
<td>&lt; 7, ≥ 7</td>
<td>90%</td>
<td>76%</td>
<td>14%</td>
<td>79%</td>
<td>80%</td>
<td>98%</td>
<td>78%</td>
</tr>
</tbody>
</table>
Table 2.2: Mayagaya et al. accuracy of mosquito age prediction (in days) when using a partial least squares regression cross-validation model (within-strain prediction) of Ifakara *Anopheles gambiae* s.s strain mosquitoes [14].

<table>
<thead>
<tr>
<th>Actual age, days</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.4</td>
<td>3.8</td>
</tr>
<tr>
<td>4</td>
<td>7.7</td>
<td>3.2</td>
</tr>
<tr>
<td>7</td>
<td>10.1</td>
<td>2.4</td>
</tr>
<tr>
<td>10</td>
<td>10.7</td>
<td>2.2</td>
</tr>
<tr>
<td>13</td>
<td>11.6</td>
<td>2.0</td>
</tr>
<tr>
<td>16</td>
<td>13.3</td>
<td>2.4</td>
</tr>
<tr>
<td>19</td>
<td>12.8</td>
<td>1.6</td>
</tr>
</tbody>
</table>

2.4 Training Our Model

We analyzed our spectra mostly as described in the Mayagaya et al. Each spectrum had an absorbance at each frequency from 350 nm to 2500 nm. Mayagaya et al. discarded all spectra with absorbance range < 0.3 and “generally outside the 0.5 - 1.0 absorbance range.” The description “generally outside” was not clear, so we reproduced it by discarding all spectra with absorbance range < 0.3 and then all spectra with any absorbance values either < 0.4, > 1.4, or both. We discarded 344 spectra out of our 786 spectra. We further trimmed noisy absorbances at frequencies from 300 nm to 499 nm and from 2001 nm to 2500 nm. We retained 442 spectra with a minimum and maximum absorbances of 0.4 and 1.4, respectively, at frequencies from 500 nm to 2000 nm.
Figure 2.2: Shows our spectra before and after cleaning. A) Spectra before cleaning (n = 786), B) Spectra after cleaning noisy ends (n = 786), C) Spectra after discarding spectra with absorbance range < 0.3 (n = 785), D) Spectra after discarding spectra with absorbances either < 0.4, > 1.4, or both (n = 442).

Figure 2.2 shows our spectra before and after cleaning. While Mayagaya et al. used PLS in Grams software PLSPlus/IQ, we used the PLS regression tool in Matlab for data cleaning, component (feature) selection, and model training and ten-fold cross-validation. We trained an actual age prediction model and interpreted results as Mayagaya et al. did: first as a regressor to estimate the actual age of mosquitoes; second as a classifier to categorize mosquitoes into three age groups.
(≤ 7d, 9-11d, or ≥ 13d); third as a classifier to group mosquitoes into two age category (either < 7d or ≥ 7d). We refer to the second and third interpretations as classifier A and B, respectively, and we use these terms in the rest of this thesis.

We further trained two classification models: the first model to classify mosquitoes into either ≤ 7d, 9-11d, or ≥ 13d age classes and the second model to classify mosquitoes into either < 7d or ≥ 7d age classes. We refer to a model trained to classify mosquitoes into either ≤ 7d, 9-11d, or ≥ 13d age classes as classification model A and the one trained to classify mosquitoes into either < 7d or ≥ 7d age classes as classification model B. The results from models A and B were compared with the results from when both Mayagaya et al., and our regression models were interpreted as classifiers.

2.4.1 The Actual Age Prediction Model

We randomised and divided processed spectra into two groups. The first group contained 332 (3/4) spectra and was used for training and cross-validation of the model. The second group had 110 (1/4) spectra and was used for out-of-sample testing. We used actual mosquito ages (1, 3, 5, 7, 9, 11, 13, 15, and 20 days old) as labels during training of the model. Figure 2.3 is the plot of regression coefficients of this model. This figure corresponds to Figure 6 in Mayagaya et al. Corresponding to the rule used by Mayagaya et al. for continuous estimation to discrete
classification, we interpreted our model output as follows: mosquitoes estimated to be younger than $< 2.5$ days were considered as 1 day old, $[2.5, 4.5)$ as 3 days old, $[4.5, 6.5)$ as 5 days old, $[6.5, 8.5)$ as 7 days old, $[8.5, 10.5)$ as 9 days old, $[10.5, 12.5)$ as 11 days old, $[14.5, 16.5)$ as 15 days old and $[19.5-21.5)$ as 20 days old. Figure 2.4 illustrates how the model fit both training and testing data. This figure is similar to Figure 5 in Mayagaya et al. There is a strong correspondence between the two figures (Figure 2.4 and Figure 5 in Mayagaya et al). From both of these figures, it is clear that the accuracies of the models (Mayagaya et al. and ours) are poor when estimating the actual mosquito ages. If the models performed well, the alignment in these figures were expected to be diagonal. The more they align horizontally, the poorer accuracy of the model. The first row in Table 2.3 shows the percentage accuracy of the model when estimating the specific age of mosquitoes.

As in Mayagaya et al., we calculated the percent accuracy by multiplying the ratio obtained by dividing the number of mosquitoes correctly estimated by the model with the total number of mosquitoes in that particular age with 100. We observed a similar pattern to that in Table 2.1 from the Mayagaya et al. While the Mayagaya et al. model had the highest percent accuracy (54%) only when predicting 10 day old mosquitoes, our model had its highest percent accuracy (33%) only when predicting 9 day old mosquitoes. For the remaining ages as in Mayagaya et al. model, our model estimation accuracy was poor. Table 2.4 shows the mean
ages and standard deviations of the model when estimating the actual age of mosquitoes. Figure 2.5 represents the five number summary results (box plot) of the same model. We observed a mean estimation close to the actual age only when the model was applied to 9 day old mosquitoes. For the other ages, the mean age estimations of the model were far from the actual age. This was consistent with Mayagaya et al. model as shown in Table 2.2. The closest mean age estimation of their model to the actual age was only when estimating ten day old mosquitoes. For the remaining ages, the mean age estimations were far from the actual ages.

When interpreting our regression model results as classifier A, the accuracy improved with only three mosquito ages (7, 9, and 11 days old) estimated with accuracies less than 50%. The remaining ages were estimated with accuracies more than 50%. This was consistent with the performance of the Mayagaya et al. model.

When interpreting our regression model as classifier B, the accuracy became even better with only two ages of mosquitoes estimated with accuracies less than 50%. This was a little different from the Mayagaya et al. model. The Mayagaya et al. model had an accuracy of less than 50% on only one age of mosquitoes when classifying them into two classes. Otherwise, we observed almost the same performance in both the Mayagaya et al. model and ours.
Figure 2.3: Regression coefficients used by our regression model to estimate ages of mosquitoes. This figure corresponds to Figure 6 in Mayagaya et al. [14].

2.4.2 Classification Models A and B

For classification model A, we divided cleaned spectra into three groups (≤ 7 days old, 9-11 days old, and ≥ 13 days old), and spectra in each group were given the same label (7 for those in ≤ 7 days old group, 10 for those in 9-11 days old group, and 13 for those in ≥ 13 days old group). We merged these groups, randomized spectra, and divided them into training (75%) and test (25%) sets.

The model was trained similarly to the actual age prediction model. Figure 2.6
Figure 2.4: Estimated ages from our regression model vs. actual mosquito ages on test data. This figure corresponds to Figure 5 in Mayagaya et al. [14].

shows the regression coefficients of this model. We interpreted the results of this model as follows: mosquitoes < 7.5 as ≤ 7 days old, [7.5, 12.5) as 9 - 11 days old and ≥ 12.5 as ≥ 13 days old.

For classification model B, we divided cleaned spectra into two groups (< 7 days old and ≥ 7 days old). The spectra in each group were labeled (6 for those in < 7 days old group and 7 for those in ≥ 7 days old group), and the two groups were merged. The spectra were randomized and divided into training (75%) and test
Figure 2.5: Box plot showing estimated ages from our regression model vs. actual mosquito ages on test data.

(25%) sets. The model was then trained using cross-validation. Figure 2.7
represents the regression coefficients of this model. During interpretation of this
model, mosquitoes < 6.5 were considered as ≤ 7 days old and ≥ 6.5 as ≥ 7 days old.
Table 2.3: Our percentage accuracy of predicting the age of *Anopheles gambiae* mosquitoes into specific age groups when using a regression model developed from the Ifakara strain of *An. gambiae* mosquitoes.

<table>
<thead>
<tr>
<th>Actual age (in days)</th>
<th>Prediction</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 7, 9-11, ≥ 13</td>
<td>Actual age</td>
<td>11%</td>
<td>8%</td>
<td>23%</td>
<td>22%</td>
<td>33%</td>
<td>8%</td>
<td>27%</td>
<td>29%</td>
</tr>
<tr>
<td>&lt; 7, ≥ 7</td>
<td>≤ 7, 9-11, ≥ 13</td>
<td>100%</td>
<td>92%</td>
<td>92%</td>
<td>33%</td>
<td>42%</td>
<td>23%</td>
<td>53%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>&lt; 7, ≥ 7</td>
<td>78%</td>
<td>38%</td>
<td>46%</td>
<td>89%</td>
<td>92%</td>
<td>77%</td>
<td>87%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2.4: Our average actual age prediction (in days) when using a partial least squares regression cross-validation model (within-strain prediction) of Ifakara *Anopheles gambiae s.s* strain mosquitoes.

<table>
<thead>
<tr>
<th>Within-strain prediction, n = 442</th>
<th>Actual age, days</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6.0</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10.9</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9.9</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9.2</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>11.9</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17.3</td>
<td>4.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5 represents the results from both classification models A and B. It is clear that the accuracy of classifying mosquitoes improved when specific classification models were trained. Classification model A had only two accuracies below 50% (only when classifying 7 and 15 day old mosquitoes into their respective age classes). The Mayagaya et al. and our regression models had three accuracies
Figure 2.6: Regression coefficients used by classification model A to classify ages of mosquitoes.

below 50% (7, 13, and 19 days old for Mayagaya et al. and 7, 9, and 11 days old for our model). Classification model B had all accuracies above 50% when classifying specific ages of mosquitoes into their respective classes. The poorest accuracy was 77%, and it was when classifying 9 day old mosquitoes. For the remaining ages, the accuracy was at least 90%. Mayagaya et al. and our regression model had accuracies below 50% when classifying 7, 3, and 5 day old mosquitoes into their
Figure 2.7: Regression coefficients used by classification model B to classify ages of mosquitoes.

Table 2.5: Our accuracy of predicting the age of *Anopheles gambiae* mosquitoes into specific age groups when using classification models developed from the Ifakara strain of *An. gambiae* mosquitoes.

<table>
<thead>
<tr>
<th>Actual age (in days)</th>
<th>Prediction</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 7, 9-11, ≥ 13</td>
<td>70%</td>
<td>54%</td>
<td>69%</td>
<td>31%</td>
<td>50%</td>
<td>57%</td>
<td>41%</td>
<td>64%</td>
<td></td>
</tr>
<tr>
<td>&lt; 7, ≥ 7</td>
<td>88%</td>
<td>100%</td>
<td>90%</td>
<td>100%</td>
<td>77%</td>
<td>92%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

respective classes. This suggests a clear improvement in accuracy when model B is used compared to when the regression model is interpreted as a classifier.
2.5 Conclusion

Our results were consistent with the results of Mayagaya et al. The highest accuracy when the Mayagaya et al. regression model estimated the actual age of mosquitoes was 24%. When the Mayagaya et al. model was interpreted as a classifier, the overall accuracies improved, especially when the output was interpreted into two age classes (< 7d or ≥ 7d). Our regression model estimated the actual age with the highest accuracy of 33%. When we interpreted it as a classifier, we obtained the highest accuracies when the output was interpreted into < 7 days old or ≥ 7 days old. Tables 2.1 and 2.3 summarize these results.

When we trained and applied classification models A and B, the accuracies were even better compared to the regression models interpreted as classifiers. Classification model A had an accuracy less than 50% only when classifying 7 and 15 day old mosquitoes, while the regression models (Mayagaya et al. and ours) had accuracies less than 50% when classifying 7, 13, and 19 day old (for Mayagaya et al. model) and 7, 9, and 11 day old (for our model). The classification model B had no accuracy below 50%. The lowest accuracy was 77% obtained when classifying 9 day old mosquitoes. The rest were at least 90%. Both Mayagaya et al. and our model had accuracies below 50% when classifying 3, 5, and 7 day old mosquitoes. (3 and 5 for our model and 7 for Mayagaya et al.), Table 2.5 summarizes this conclusion.
We conclude reproducibility of Mayagaya et al. study and emphasize application of NIRS technology for estimating the age of mosquitoes. We recommend training and application of classification model over regression model to classify age of mosquitoes into $< 7$ or $\geq 7$ days old.
CHAPTER 3

Do NIR Spectra Collected from Lab-reared Mosquitoes Differ from Those Collected from Wild Mosquitoes?

This chapter presents cluster analysis performed on a mixture of spectra collected from lab-reared and wild mosquitoes of the same species to determine if there is any significant difference between these two groups of spectra. This chapter is a preliminary version of [15] in preparation for submission to a peer-reviewed journal.

3.1 Mosquito Age and Malaria Transmission

Female mosquitoes in the genus Anopheles are known to be threat to public health as they transmit malaria, a parasitic disease which killed approximately 438,000 people in 2015 [35]. Mosquitoes contribute to malaria transmission by hosting and allowing the development to maturity of the malaria-causing Plasmodium parasite. Mosquitoes acquire Plasmodium when they feed on infected human blood for their egg development. Once in the mosquito depending on temperature, Plasmodium takes 10-14 days to develop fully enough to cause malaria in a human [3]. Hence, if a mosquito is less than 10 days old, the chances that it is carrying a full developed parasite (infectious) is small [3]. Therefore, knowing the age of a mosquito is crucial
in evaluating the infectiousness of the Anopheles mosquito population in an area, especially when vector control resources are limited.

Currently, a skilled lab technician estimates the age of a mosquito by hand dissection of her ovaries to determine whether she has laid eggs. Those found to have laid eggs are assumed older than those who have not laid eggs [7]. Because hand dissection is laborious and difficult, its application is limited to only few experts working with small numbers of mosquitoes. In addition, this method can classify mosquitoes as relatively young (not laid eggs) or relatively old (laid eggs), but it cannot infer mosquito age in days. As a result, we need a new inexpensive mosquito age determination approach.

Near Infrared Spectrometry (NIRS) is an alternative to hand dissection. It is a high throughput, automated technique, which measures the amount of the near infrared energy absorbed by samples. NIRS has been applied to identify species of insects infecting stored grains [9]; to age grade houseflies [20], stored-grain pests [21], and biting midges [22]; to differentiate between species and subspecies of termites [1]; to estimate the age and identify species of morphologically indistinguishable laboratory reared and semi-field raised Anopheles gambiae and Anopheles arabiensis [14, 26]; to detect and identify two strains of Wolbachia pipiens (wMelPop and wMel) in male and female laboratory-reared Aedes aegypti mosquitoes [28]; and to classify the age of male and female wild-type and Wolbachia
infected *Aedes aegypti* [27]. The study by Mayagaya et al. reports that NIRS can estimate the age of lab-reared and semi-field mosquitoes into either less than 7 or greater than 7 days old with an accuracy exceeding 80% [14].

The ability of NIRS to estimate the age of laboratory and semi-field raised mosquitoes is not what is needed. Lab-reared mosquitoes do not transmit malaria; wild mosquitoes do. It is unknown if NIRS can estimate the age of wild mosquitoes because we lack age-labeled wild mosquitoes with which to train machine learning algorithms. We can apply a model trained on spectra from lab-reared mosquitoes to estimate the age of wild mosquitoes, but no studies have validated that generalization, which would be appropriate only if spectra collected from lab-reared mosquitoes are equivalent to those collected from wild mosquitoes.

While we cannot validate the generalization of an age classification model from lab-reared to wild mosquitoes, here we explore whether spectra from lab-reared mosquitoes differ from spectra from wild mosquitoes. We performed $k$-means cluster analysis: first, on a mixture of spectra from different ages (1, 3, 5, 7, 9, 11, 15, 20, and 25 days old) of lab-reared and from all wild collected *Anopheles arabiensis*; second, on a mixture of spectra from lab-reared, and wild *Anopheles arabiensis* while controlling the age of lab-reared mosquitoes; and third, on the mixture of spectra with their features reduced to ten components after applying partial least squares (PLS) on mixed (lab-reared and wild) raw spectra. At the end of each
cluster analysis, we performed a $X^2$ statistical test to determine the correlation between the formed clusters and the type of spectra. Our null hypothesis is: There is no significant difference between the spectra collected from lab-reared and those from wild mosquitoes, when other factors are equal. We accept the null hypothesis based on the results from all the three approaches outlined above.

3.2 Overview of Cluster Analysis

Cluster analysis is an unsupervised data partitioning process that groups a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some way) to each other than to those in other groups (clusters). The term “unsupervised” means that during cluster analysis, no labels are given to the objects; clustering depends only on the set of features describing each object [33]. If we translate this definition to our problem, it means that during analysis, we do not label spectra as lab or wild, we only provide spectra features (absorbances). If spectra collected from lab-reared and those from wild are different, we expect them to be grouped (clustered) in different clusters, otherwise they should distribute equally in the formed clusters.

In this study, we applied $k$-means cluster analysis. $K$-means cluster analysis, also known as Lloyd’s algorithm [12], starts by arbitrary picking cluster centers known as centroids, depending on the number of clusters needed. In our case, we
need two clusters, so the number of centroids is two. The next step is to compute distances from each object (spectrum in our case) to each centroid and assign each object to its closest centroid. There are different ways to compute distance, but this algorithm uses squared Euclidean distance [36]. The average distance of objects assigned to each centroid is then computed. The process repeats by selecting new centroids and reassigning objects until the average distance to centroids is minimized. More about \( k \)-means clustering can be found in [2, 12].

When the clusters are formed, the next step is to evaluate their quality. One measure of cluster quality is the silhouette coefficient (SC) [24]. SC shows how clusters are separated from one another and how tight objects are in each cluster. The SC, which is computed using Equation 3.1, ranges from \(-1\) to \(+1\). A high SC indicates that an object is well-matched to objects in its own cluster and poorly-matched to objects in neighboring clusters. If most objects in the cluster have high SC, then the clustering is appropriate. If many objects in the cluster have a low or negative SC, then the clustering solution is inappropriate. SC of the cluster is an average of all SC of objects in that cluster. Table 3.1 summarizes the interpretation of average SC of objects in a cluster [29].
Table 3.1: Interpretation of the silhouette values for partitioning methods [29].

<table>
<thead>
<tr>
<th>Silhouette coefficient</th>
<th>Proposed Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71-1.00</td>
<td>A strong cluster has been found</td>
</tr>
<tr>
<td>0.51-0.70</td>
<td>A reasonable cluster has been found</td>
</tr>
<tr>
<td>0.26-0.50</td>
<td>The cluster is weak and could be artificial</td>
</tr>
<tr>
<td>≤ 0.25</td>
<td>No substantial cluster has been found</td>
</tr>
</tbody>
</table>

Let

\[ s(o) = \text{Silhouette coefficient of a single object ‘o’} \]

\[ a(o) = \text{Average distance of object ‘o’ to the other objects in its cluster} \]

\[ b(o) = \text{Average distance of object ‘o’ to other objects in the nearest cluster}. \]

Then

\[ s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}. \] (3.1)

Note: The lower the \( a \) value the better, and the higher the \( b \) value the better.

### 3.3 Mosquitoes and Spectra Collection

We used lab-reared *Anopheles arabiensis* mosquitoes with ages 1, 3, 5, 7, 9, 11, 15, 20, and 25 days from the Ifakara Health Institute insectary. This insectary started in 2010 and is reared in a semi-field system (SFS) [17] under ambient temperature and light-dark cycles. The humidity is artificially increased to 80% during the dry season (May - October). Adult mosquitoes are daily provided a human arm as a
blood meal source and 10% glucose solution. The insectary keeps records of mosquitoes from egg laying to adult emergence, and the cages are labeled in such a way mosquito ages are easily identified.

Wild *Anopheles arabiensis* mosquitoes were collected using CDC light traps [30] in Minepa, a village located in south-eastern Tanzania. The traps were set in selected houses in the evening and collected in the next morning. Live *Anopheles gambiae* complex were sorted from the traps and put in a small cage with cotton dipped in 10% sugar solution at the top of the cage. The sorted live *Anopheles gambiae* complex were transported to Ifakara Health Institute laboratory for spectra collection (scanning).

Before scanning, both lab-reared and wild mosquitoes were killed by freezing for 20 minutes. We collected spectra using the machine and followed procedures as described in Mayagaya et al. [14]. After scanning, wild mosquitoes were dissected to determine their egg laying status followed by polymerase chain reaction (PCR) for species identification [19]. Using PCR results, we sorted wild *Anopheles arabiensis* spectra for analysis.
3.4  \textit{K}-means Clustering on a Mixed Spectra

After spectra collection, we removed associated spectra labels and performed \textit{k}-means cluster analysis in three different ways: first, we mixed 871 spectra collected from all ages (i.e., 1, 3, 5, 7, 9, 11, 15, 20, and 25) of lab-reared \textit{Arabiensis} and 947 spectra collected from wild \textit{Arabiensis} and performed \textit{k}-means cluster analysis using the cluster analysis tool in Matlab. Figure 3.1 shows a two-dimensional plot of clusters displayed using absorbances at 350nm and 351nm. We generated similar displays using absorbances at different frequencies, and the patterns of the displays were similar. Figure 3.1 shows that there are two clusters, despite some overlapping of spectra (objects) in both clusters.
Figure 3.1: Two-dimensional plot of clusters using absorbances at 350 nm and 351 nm, when all ages of lab-reared mosquitoes were used in the analysis.

To evaluate the quality of our clusters, we generated Figures 3.2 and 3.3, which represent silhouette coefficients (SC) of each spectrum (object) in its cluster. The figures show that most spectra (objects) have an acceptable SC, and the average SC of each cluster (0.77 and 0.53 for clusters one and two, respectively; refer to Table 3.1 for an interpretation of average SC) shows that the clusters visualized in Figure 3.1 are both strong and reasonable, respectively.
Figure 3.2: Silhouette coefficient of each spectrum (object) in its associated cluster when we mixed all ages of lab-reared with wild *Anopheles Arabiensis*. The average SC for cluster one and two is 0.77 and 0.53 indicating the clusters are strong and reasonable, respectively.

Assured from Table 3.1 that our clusters are strong, we generated a contingency table and performed a $X^2$ statistical test to determine if there is a significant difference in distribution of lab-reared and wild mosquitoes in our two clusters. That is, do the two clusters capture the sources of the mosquitoes? Table 3.2 summarizes the results, showing unequal distribution of both lab-reared and
Figure 3.3: Silhouette coefficients of each spectrum (object) in its associated cluster when we mixed all ages of lab-reared with wild *Anopheles Arabiensis*.

Wild mosquitoes in the clusters. Cluster one has more lab-reared mosquitoes, while cluster two has more wild mosquitoes.

What does $X^2$ say? It suggests that there may be a significant difference between the spectra of wild and lab-reared mosquitoes. However, we suspect the clustering might be age-related, not source-related, and a 2016 report from the American Statistical Association strongly warns against narrow interpretations of
Table 3.2: Distribution of mosquitoes in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c.X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$), and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise) when all ages of lab-reared mosquitoes were included in the analysis.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Lab</th>
<th>Wild</th>
<th>Total</th>
<th>Av. SC</th>
<th>c. $X^2$</th>
<th>$P_{c.X^2}$</th>
<th>$X^2_{\alpha=0.05}$</th>
<th>Assoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>510</td>
<td>488</td>
<td>998</td>
<td>0.77</td>
<td>2</td>
<td>0.003</td>
<td>3.84</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>361</td>
<td>459</td>
<td>820</td>
<td>0.53</td>
<td>9.04</td>
<td>0.003</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>871</td>
<td>947</td>
<td>1818</td>
<td>0.53</td>
<td>9.04</td>
<td>0.003</td>
<td>3.84</td>
<td></td>
</tr>
</tbody>
</table>

p-values [34] in favor of critical assessment of an entire statistical model. Hence, we explore possible age-dependencies that may influence our clustering.

We repeated the $k$-means analysis, this time controlling the ages of lab-reared mosquitoes. We could not control the ages of wild mosquitoes as they are unknown. The purpose of repeating the analysis while controlling the age was to test whether the results we got in the first approach may have been influenced by age. We mixed 80 raw spectra (without any cleaning) of one day old lab-reared *Arabiensis* and 80 randomly selected spectra from wild *Arabiensis* and performed the analysis as in the first approach. We repeated the process for the remaining ages (i.e., 3, 5, 7, 9, 11, 15, 20, and 25) of lab-reared mosquitoes, while keeping the spectra from wild *Arabiensis* unchanged (same 80 randomly selected). Table 3.3 summarizes the results, showing that the lab-reared and wild spectra were not correlated (independent) only when clustering involved 3, 5, and 25 day old mosquitoes.
lab-reared mosquitoes. For the other ages, the spectra were correlated (dependent). The outcome strengthen our hypothesis on age influencing the previous clustering.

These results lead to the next approach. We removed spectra associated with 3, 5, and 25 day old lab-reared mosquitoes from the lab-reared data set. We then mixed with all the wild spectra and performed the analysis as in the first approach. Figure 3.4 shows a two-dimensional plot of the clusters displayed using absorbances at 350 and 351 nm. Figure 3.4 shows no obvious change in the display of clusters when compared with Figure 3.1 from the first approach. Figures 3.5 and 3.6 show the silhouette coefficients of objects in each cluster. These figures show the same pattern of SC as Figures 3.2 and 3.3 in the first approach, which means the quality of clusters were not compromised with the removal of 3, 5, and 25 day old lab-reared mosquitoes in the analysis. Table 3.4 represents the results, which show the association between spectra collected from lab-reared and wild mosquitoes of the same species. This suggest that the results in the first approach (Table 3.2) were influenced with mosquito age differences and not their source (lab-reared or wild).

In our last approach, we performed partial least square (PLS) on spectra to reduce data dimension and used ten PLS components to perform $k$-means cluster analysis as in the first approach. Figures 3.7, 3.8, and 3.9 show the two dimensional display and SC of the formed clusters. Using interpretation on SC summarized in Table 3.1, it is clear from these figures that no substantial clusters were formed.
This strengthens the results we obtained in the third approach (when spectra from 3, 5, 25 days old lab-reared mosquitoes were not included in the cluster analysis), where we found no significant difference ($p = 0.245$) between spectra collected from lab-reared and wild mosquitoes of the same species. The results further suggest that the chances that clustering in the first approach was influenced by age, is high.

Figure 3.4: Two dimensional plot of clusters using absorbances at 350nm and 351nm when 3, 5, and 25 day old lab-reared mosquitoes were removed during cluster analysis.
Figure 3.5: Silhouette coefficient of each spectrum (object) in its associated cluster when 3, 5, and 25 day old lab-reared mosquitoes were not included in the analysis.

3.5 Discussion - What do results from this study mean?

Our results show that cluster analysis on the mixture of spectra without controlling the age of lab-reared mosquitoes produced clusters associated with the source of the spectra. This could infer that there is difference in spectra collected from lab-reared mosquitoes to that collected from the wild. However, as the American Statistical Association (ASA) cautioned on p-value interpretation [34], different factors apart from the source of the spectra may have contributed to the results. Age of a
mosquito is one of the most important factors to consider, as different studies [14, 26] have already shown that spectra can be used to estimate the ages of mosquitoes, implying that mosquitoes of the same species but different ages can be differentiated using spectra. Hence, clustering of spectra can occur based on age differences of mosquitoes. Physiological status (laid eggs or not, blood fed or not) of a mosquito can also influence the cluster formation. Ntamatungiro et al [18] showed there is an influence of physiological status of a mosquito on the spectra. Also from

Figure 3.6: Silhouette coefficient of each spectrum (object) in its associated cluster when 3, 5, and 25 day old lab-reared mosquitoes were removed during cluster analysis.
the on-going discussion about $p$-values [10], it is argued that $p$-values are associated with the statistical analysis or model used during the study. Therefore, a decision to accept or reject the null hypothesis should be based not only on a $p$-value, but also must consider other factors, such as the impact of age, in our case.

Therefore, we explored whether the age of mosquitoes might be influencing the results in the first approach. We repeated the cluster analysis on the mixture of spectra, while controlling the age of lab-reared mosquitoes. We could not control
the age of wild mosquitoes, as we did not have age labels for wild mosquitoes.

Instead, we randomly selected 80 spectra collected from wild mosquitoes and
maintained it for the rest of the analysis, while changing the age of the lab-reared
mosquitoes. The results showed association between clusters and sources of the
spectra only when we performed cluster analysis with 3, 5, and 25 day old
lab-reared mosquitoes. This means that there is a possibility that in the original
wild data set, the presence of very young mosquitoes might have influenced the
Figure 3.9: Silhouette coefficients of each spectrum (object) in its associated cluster when cluster analysis uses ten PLS components.

results. When we removed spectra from 3, 5, and 25 day old mosquitoes from the original lab-reared data set and repeated cluster analysis using all spectra as in the first approach, the clusters did not associate with the source of the spectra, meaning that the chance that age influenced the results in the first approach is high. When we performed cluster analysis while controlling the egg laying status (as one way to determine the influence of physiological status) of both wild and lab-reared mosquitoes, results showed no influence on cluster formation.
We further performed the partial least square on the spectra to reduce features before we did cluster analysis. Feature reduction using PLS can help during analysis as it reduces noisy in data without loosing important information. Initially, the spectra had 2152 features, which can introduce errors during cluster analysis. Partial least square (PLS) does not discard any information when reducing features; instead it finds components associated with all features while considering dependent variables [23, 25]. When we applied PLS and performed clustering on the reduced features (ten components), we found very poor clustering, with average SC below 0.21, which indicates that there is no clustering tendency in the data [24, 29].

These results strengthened the results obtained when the age of lab-reared mosquitoes was controlled. Based on these facts and arguments, we fail to reject the null hypothesis.
3.6 Conclusion - No Clear Difference Between Spectra Collected from Lab-reared and Wild Mosquitoes of the Same Species

Our study concludes that there is no clear difference between spectra collected from lab-reared and wild mosquitoes. When the age of lab-reared mosquitoes was controlled during the analysis, we found no significant difference in the distribution of lab-reared and wild mosquitoes between the two clusters and failed to reject the null hypothesis \( P = 0.245 \). The clusters were represented with average silhouette values of 0.51 (reasonable) and 0.77 (strong), respectively [29].

These results strengthen the idea of training models to estimate the age of wild mosquitoes using spectra collected from lab-reared mosquitoes. The accuracy of NIRS on estimating the age of wild mosquitoes can be assumed to be similar to that on lab-reared mosquitoes.
Table 3.3: Distribution of spectra in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c,X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$) and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise), when the ages of lab-reared were controlled before clustering.

<table>
<thead>
<tr>
<th>Age</th>
<th>Cluster</th>
<th>Lab</th>
<th>Wild</th>
<th>Total</th>
<th>Av. SC</th>
<th>c. $X^2$</th>
<th>$P_{c,X^2}$</th>
<th>$X^2_{\alpha=0.05}$</th>
<th>Assoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>34</td>
<td>38</td>
<td>72</td>
<td>0.48</td>
<td>0.77</td>
<td>0.40</td>
<td>0.53</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>46</td>
<td>42</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>33</td>
<td>46</td>
<td>79</td>
<td>0.64</td>
<td>0.65</td>
<td>4.23</td>
<td>0.04</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>47</td>
<td>34</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>46</td>
<td>29</td>
<td>75</td>
<td>0.69</td>
<td>0.69</td>
<td>7.31</td>
<td>0.01</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>34</td>
<td>51</td>
<td>85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>47</td>
<td>38</td>
<td>85</td>
<td>0.67</td>
<td>0.73</td>
<td>2.03</td>
<td>0.15</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>42</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>37</td>
<td>42</td>
<td>79</td>
<td>0.71</td>
<td>0.66</td>
<td>0.63</td>
<td>0.43</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>43</td>
<td>38</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>30</td>
<td>40</td>
<td>70</td>
<td>0.82</td>
<td>0.41</td>
<td>2.54</td>
<td>0.11</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>50</td>
<td>40</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>34</td>
<td>43</td>
<td>77</td>
<td>0.45</td>
<td>0.78</td>
<td>2.03</td>
<td>0.15</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>46</td>
<td>37</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>35</td>
<td>42</td>
<td>77</td>
<td>0.60</td>
<td>0.74</td>
<td>1.23</td>
<td>0.27</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>45</td>
<td>38</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>47</td>
<td>29</td>
<td>76</td>
<td>0.66</td>
<td>0.70</td>
<td>8.12</td>
<td>0.004</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>51</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>80</td>
<td>80</td>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4: Distribution of spectra in clusters, average SC (Av. SC) of each cluster, computed $X^2$ (c. $X^2$), p-value of the computed $X^2$ ($P_{c,X^2}$), $X^2$ at $\alpha = 0.05$ ($X^2_{\alpha=0.05}$) and association (Assoc) between formed clusters and the type of spectra (yes if there is association, no if otherwise), when 3, 5, and 25 day old lab-reared mosquitoes were not included in the analysis.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Lab</th>
<th>Wild</th>
<th>Total</th>
<th>Av. SC</th>
<th>c. $X^2$</th>
<th>$P_{c,X^2}$</th>
<th>$X^2_{\alpha=0.05}$</th>
<th>Assoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>335</td>
<td>501</td>
<td>836</td>
<td>0.51</td>
<td>0.77</td>
<td>0.245</td>
<td>3.84</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>264</td>
<td>446</td>
<td>710</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>599</td>
<td>947</td>
<td>1546</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 4

Age Classification of Wild Mosquitoes Using a Classification Model Trained on Lab-reared Mosquitoes (Model B in Chapter Two)

This chapter presents the results obtained when a classification model trained on lab-reared Anopheles arabiensis mosquitoes is applied to classify the age of wild Anopheles arabiensis.

4.1 Age-class Estimation of Wild Anopheles arabiensis Using Model B Developed from Lab-reared Anopheles arabiensis in Chapter Two

Evaluation of malaria control interventions such as the use of insecticide-treated nets (ITNs) often focuses on assessing changes in the mosquito age structure. A shift in the age structure towards a younger non-infectious population signifies a reduction in the risk of malaria transmission. The current gold standard for mosquito age-grading involves dissection of ovaries to separate mosquitoes into those that have previously laid eggs (parous) and those that have not laid eggs (nulliparous). However, this technique is time consuming and laborious. Near infrared spectrometry (NIRS) can be an alternative to ovary dissection; it is instantaneous and does not require reagents, so hundreds of samples can be scanned in a day. Mayagaya et al. 2009 demonstrated near-infrared
spectroscopy (NIRS) can be at least 80% accurate for classifying the age of lab-reared mosquitoes [14]. In Chapter Two of this thesis, using a different set of lab-reared mosquitoes, we successfully reproduced the Mayagaya et al. accuracy result.

Despite of an ability of near-infrared spectrometry to classify the age of lab-reared mosquitoes, it is unknown if NIRS can classify the age of wild mosquitoes. The limitation is the lack of age-labeled wild mosquitoes with which to train the model. Training a model using labels from ovary dissection yields a model with poor accuracy. Applying a model trained on spectra from lab-reared mosquitoes to estimate the age of wild mosquitoes would be an appropriate only if spectra collected from lab-reared mosquitoes are equivalent to those collected from wild mosquitoes, but no studies have validated that generalization.

In Chapter Three, we applied $k$-means cluster analysis to a mixture of spectra collected from lab-reared and wild *Anopheles arabiensis* mosquitoes to determine if there is a significant difference between these spectra. We found no clear difference in the distribution of both wild and lab-reared mosquitoes between the two formed clusters. The clusters had average silhouette coefficients (a measure of cluster quality) of 0.51 and 0.77, respectively, which suggests that the clusters are reasonable and strong, respectively.

Knowing from Chapter Three that there is no significant difference between
spectra collected from lab-reared and wild mosquitoes of the same species, in this chapter, we applied a classification model B developed in Chapter Two to estimate age classes of wild *Anopheles arabiensis* mosquitoes. These wild mosquitoes were collected from March to October using CDC light traps in Minepa, a village in south-eastern Tanzania. Only live *Anopheles gambiae* complex were sorted and transported to the Ifakara Health Institute (IHI) laboratory for spectra collection, ovary dissection (to determine their egg laying status), and polymerase chain reaction (PCR) analysis to identify *Anopheles arabiensis* mosquitoes. Figure 4.1 summarizes the process from collection of wild mosquitoes to spectra collection, ovary dissection, and species identification.

![Figure 4.1: Process of wild mosquito collection, spectra collection, ovary dissection, and species identification.](image)
Next in this chapter, we describe how spectra were collected from the wild mosquitoes; how mosquito ovaries were dissected to get their egg-laying status; and how wild *Anopheles arabiensis* mosquitoes were identified from other morphological indistinguishable wild mosquitoes for the model B to classify the age. We then present results obtained and compare them with egg-laying classifications from Detinova ovary dissections and with historical studies [4, 8, 32] conducted to determine the age structure of wild mosquito population. Finally, we draw conclusions about the accuracy of our model by comparing the number of mosquitoes in each age class obtained when model B and Detinova ovary dissection are applied independently to classify the age of the same set of wild mosquitoes and how the model age classification relates to historical studies on the age structure of wild mosquito populations.

4.2 Spectra Collection and Ovary Dissection of Wild *An. arabiensis*

At the IHI laboratory, live wild mosquitoes received from Minepa village were killed by freezing before scanning them using a LabSpec 5000 near-infrared spectrometer (ASD Inc, Boulder, CO) to collect spectra. Figure 4.2 shows the spectra collection process. The details on how spectra collection was conducted can be found in Chapter Two or in [14].

After collecting its spectrum, each mosquito was labelled with the spectral
number obtained during scanning. This unique identifier was used throughout the processing of that mosquito to link its absorbance information with its egg laying status (obtained after ovary dissection) and species information obtained after PCR. Ovary dissection shown in Figure 4.3 was performed according to Detinova [6] soon after NIR spectra collection. PCR for species identification was run on mosquito legs after ovary dissection.
Figure 4.3: Ovary dissection of wild *Anopheles gambiae* complex mosquitoes before their species identification by PCR. A) Dissected mosquitoes; B) Mosquito Dissection process; C) Zoomed in dissection process.

4.3 Age Classification of Wild *Anopheles arabiensis*

Using mosquito species labels obtained after PCR, required species of wild mosquitoes (*Anopheles arabiensis*) with known egg laying status were sorted. Our classification model B, which was trained on lab-reared *Anopheles arabiensis* in Chapter Two of this thesis, was applied to classify their age as young (< 7 days old) or old (≥ 7 days old). As shown in Table 4.1, the model classified a large number of mosquitoes as young and a small number as old.
Table 4.1: Age classification of wild *Anopheles arabiensis* by model B.

<table>
<thead>
<tr>
<th>Model age classification</th>
<th>Number of mosquitoes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young (&lt; 7 days old)</td>
<td>797 (86%)</td>
</tr>
<tr>
<td>Old (≥ 7 days old)</td>
<td>130 (14%)</td>
</tr>
<tr>
<td>Total</td>
<td>927 (100%)</td>
</tr>
</tbody>
</table>

Because we lack age labels of wild *Anopheles arabiensis*, we cannot directly validate the accuracy of the model classification presented in Table 4.1.

Alternatively, the age classification by model B was validated indirectly:

1. By analyzing the distribution of nulliparous (not laid eggs) and parous (laid eggs) mosquitoes in each age-class from the model;

2. By comparing the number of mosquitoes in each age class obtained when classification is done using model B and when done using Detinova ovary dissection; and

3. By relating with the historical studies conducted to determine the age structure of wild mosquito populations.

We now consider each in turn.
4.4 Analysis on the Number of Nulliparous and Parous Mosquitoes in Each Age-class Obtained from Model B

Normally for a mosquito to lay eggs, she has to mate, feed on blood, and rest for at least three days post blood feeding for digestion and egg development [13]. Based on this egg laying cycle, the chance that a wild mosquito has laid eggs when she is less than seven days old is lower than when she is seven or more days old. The classification model B trained in Chapter Two of this thesis classifies mosquitoes as either less than seven days (young) or more than or equal to seven days (old). Therefore, we expect to see a correlation between model age classification and mosquito egg-laying status, where most mosquitoes which have laid eggs should be classified by model B as at least seven days old, and most of those which have not laid eggs should be classified as less than seven days old.

We counted the number of parous and nulliparous mosquitoes in each age-class obtained from model B and performed Pearson’s $X^2$ analysis to determine if there is association between model age classification and mosquito’s egg-laying (parity) status. Based on mosquito egg laying cycle, our hypothesis is that there is a correlation between model age classification and mosquito egg laying status.

As presented in Figure 4.4, our model B classified most of both parous and nulliparous mosquitoes as less than seven days old, showing no association ($X^2 = 1.98, \text{ and } X^2_{a=0.05} = 3.84$) between model age classification and mosquito egg
Figure 4.4: Percentage distribution of nulliparous and parous mosquitoes in each age-class from model B.

laying status. Despite of these results, it is still difficult to rule out any correlation between model age classification and mosquito parity status. Probably, most of these laid-eggs mosquitoes, which the model estimated to be less than seven days old, might have laid eggs once (one gonotrophic cycle). From the mosquito egg-laying cycle, it is possible for a mosquito to lay eggs once, and almost impossible to lay eggs twice when she is less than seven days old [13]. If we knew how many times each mosquito has laid eggs, we could infer any correlation between
model age classification and mosquito egg laying status. The Detinova method used in this study does not tell how many times a mosquito has laid eggs; it only tells whether she has laid eggs. Polovodova [4], another type of mosquito ovary dissection to estimate the age-class of mosquitoes, estimates how many times a mosquito has laid eggs. This method can be a candidate for future studies to estimate any correlation between model age classification and mosquito egg laying status. However, the Polovodova method is more cumbersome than Detinova, and even fewer experts can do it.

As shown by Mayagaya et al. [14] and by us in Chapter Two, re-training the model by changing the model cutoff point to an age below or above seven days old compromises the accuracy of the model. No study has explored why, but perhaps at seven days old, a substantial chemical composition change might be happening in Anophelines mosquitoes, which affects their light absorbance enough to be picked up by the model during training. Possibly, this change is not pronounced in other age stages of mosquitoes. Normally, chemical composition changes must reach a level of parts-per-thousand or greater to be detected by NIRS [14] during training.

Failing to identify any correlation between model age classification and mosquito egg laying status, in the next section, we compared the number of mosquitoes in each age class obtained when model B and Detinova ovary dissection are independently applied to classify the age of mosquitoes.
4.5 Comparison Between Number of Mosquitoes in Each Age Class Obtained when Model B and Detinova Ovary Dissection Are Separately Used to Classify Age of the Same Set of Mosquitoes

After analyzing the distribution of parous and nulliparous in each age-class obtained from model B and failing to identify any correlation between model age classification and mosquito egg-laying status, in this section, we compare the number of mosquitoes in each age-class obtained when the model and Detinova ovary dissection applied separately to classify the age of same set of mosquitoes. We assumed no correlation between model age classification and mosquito egg laying status, and we ignore the distribution of nulliparous (not laid) and parous (laid eggs) mosquitoes in age-classes formed from the model.

As shown in Figure 4.5, our model B classified 86% of the total 927 mosquitoes as young (less than seven days old) and 14% as old (greater or equal to seven days old). Detinova ovary dissection classified 72% of the same number (927) of mosquitoes as young (not laid eggs) and 28% as old (laid eggs). Using information from Table 4.2 and Figure 4.4, we applied Jaccard similarity coefficient [11, 31] to determine similarity between the outputs of the two methods (Detinova ovary dissection and model B) and found it to be 67% (Meaning there is 67% chances that the two methods will classify a mosquito into the same age class
Figure 4.5: Percentage of mosquitoes in each age class of wild *Anopheles arabiensis* obtained when model B and Detinova ovary dissection are separately applied to classify the age of the same population of mosquitoes (N = 927).

and 33% in different age class). Hence, model B and Detinova ovary dissection are more similar than they differ.

We further split wild mosquitoes into the month of collection and repeat the analysis (compare the number of mosquitoes in each age class obtained when model B and Detinova ovary dissection are separately applied to classify the age of the same set of mosquitoes). Figure 4.6 shows the results obtained. Despite observable
Table 4.2: The number of mosquitoes in each age class of wild mosquitoes obtained when model B and Detinova ovary dissection are independently applied to classify the age of the same population of mosquitoes.

<table>
<thead>
<tr>
<th></th>
<th>Model B</th>
<th>Ovary dissection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young</td>
<td>Old</td>
</tr>
<tr>
<td>Young</td>
<td>581</td>
<td>87</td>
</tr>
<tr>
<td>Old</td>
<td>216</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>797</td>
<td>130</td>
</tr>
</tbody>
</table>

deviations in June and in October, there is agreement between the age classifications from the two methods.

We cannot explain why we find such agreement between the two methods when we separately compare them, while we find no association between model age classification and mosquito’s egg-laying status.

These results suggest that model B can be a reliable method to complement the Detinova ovary dissection method. Since NIRS is a high-throughput technique, complementing Detinova ovary dissection with model B allows entomologists and other stakeholders to draw conclusions about the age composition of a particular wild mosquito population based on a statistically acceptable sample size. Currently, because of the tediousness of ovary dissection, entomologists infer the age distribution of a wild mosquito population based on small sample sizes, which statistically might not represent the true population.
Figure 4.6: Comparison on the percentage of mosquitoes in each age class obtained when model B and Detinova ovary dissection are independently applied to classify the age of the same set of wild Anopheles arabiensis per month of collection.

4.6 Comparison Between Model B Age Classification and Other Studies on Age Structure of Wild Mosquitoes

As already shown in Table 4.1, model B classified 86% of wild Anopheles arabiensis as young and 14% as old. These results agree with other studies [4, 5, 8, 32] conducted to estimate the age structure of wild mosquitoes.

Figure 4.7, which was modified from one of the studies [5], summarizes the conclusions from all the studies [4, 5, 8, 32], showing that wild mosquito populations
generally contain more young (not laid eggs or laid eggs once) mosquitoes than old
(laid eggs more than once) mosquitoes.

Figure 4.7: Modified from Brownstein et al. [5] showing age structure of a simulated
wild mosquito population. Yellow bars represent mosquitoes which are less likely to
be infectious, and red bars represent potential infectious mosquitoes.
The studies [4, 8] specifically point out that most of mosquito samples collected during the rainy season include more young than old mosquitoes, and those collected in the dry season consist of more old than young mosquitoes because during the rainy season, the rate of mosquito emergence is higher than the mosquito survival rate, and during the dry season, the emergence rate decreases to below the survival rate. The wild *Anopheles arabiensis* used in our study were collected in the Kilombero valley in south-eastern Tanzania. The valley often has heavy rains from January to May, followed by light rains for June and July. This prolongs the rainy season behavior (stagnant water bodies), which supports mosquito breeding and a high mosquito emergence rate, even after the rainy season has ended. We collected a large portion of mosquitoes from March to August, with a few collected in September and October. According to the specific findings of [4, 8], our mosquitoes are expected to consists more young than old mosquitoes, which supports results from our model.

Therefore, the knowledge from historical studies on the age structure of wild mosquitoes supports the age classification results obtained from the model B.
4.7 Model B Trained on Lab-reared Mosquitoes Correctly Classifies the Age of Wild Mosquitoes

In conclusion, we find our model B trained on NIR spectra collected from lab-reared mosquitoes can be relied on to classifying the age of wild mosquitoes of the same species. Our conclusion is based on the following findings: 1. The age classification results of wild mosquitoes obtained when model B is used to classify a particular set of wild mosquitoes is consistent with the age classification results obtained when Detinova ovary dissection is separately applied to classify the same set of wild mosquitoes. 2. The age classification results obtained from the model B agree with the results from other studies performed to estimate the age structure of wild mosquito populations [4, 5, 8, 32]. 3. The accuracy of our model B in classifying the age of lab-reared *Anopheles arabiensis* is greater than 80% [14], and our study in Chapter Three has shown that there is no significant difference between spectra collected from lab-reared and wild *Anopheles arabiensis*.

We recommend a replica study using the Polovodova ovary dissection instead of Detinova ovary dissection to clear any remained doubts on the accuracy of our model B to classify the age of wild mosquitoes.
CHAPTER 5

No Clear Difference on Spectra Collected from Lab-reared and Wild Mosquitoes: A Classification Model Trained on Spectra from Lab-reared Mosquitoes Classifies Wild Mosquitoes Consistently with Classification by Detinova Ovary Dissection

This chapter summarizes the findings and discusses our contributions to the fight against malaria. It also provides suggestions and recommendations for further studies.

5.1 Summary of Thesis Findings and Contributions to the Fight Against Malaria

In Chapter Two, the thesis replicates the published [14] accuracy of near infrared spectrometry for classifying the age of lab-reared mosquitoes. In Chapter Three, we find no clear difference between spectra collected from lab-reared and wild mosquitoes of the same species, which suggests that applying a model trained on spectra from lab-reared mosquitoes to estimate the age of wild mosquitoes can be appropriate. In Chapter Four, a classification model trained on lab-reared mosquitoes classifies the age of wild mosquitoes consistently with age classification obtained when Detinova ovary dissection [7] is applied to classify the same set of
mosquitoes. The model age classification also agrees with earlier studies [4, 5, 8, 32] conducted to estimate the age structure of wild mosquitoes.

Our results strongly suggest the reliability of a NIRS model for classifying wild mosquitoes as young (less than 7 days old) or old (≥ 7 days old), as a complement to Detinova ovary dissection. Our results add valuable information to previously established information on NIRS as a complementary method to classify age of wild mosquitoes, allowing entomologists and other stakeholders to make more informed decisions on the use of the method.

Using a NIRS classification model as a complementary method to Detinova ovary dissection when classifying ages of wild mosquitoes has the following advantages:

Provides entomologists a large and statistically acceptable sample size to infer a population age profile. Currently, the tediousness and time-consuming nature of the ovary dissection method limits entomologists to small sample sizes. With near infrared spectrometry, entomologists can scan hundreds of mosquitoes per day, against fewer than fifty mosquitoes with ovary dissection.

Using NIRS to estimate age is time and cost-effective compared to ovary dissection. No reagents are required when using NIRS to estimate the age of mosquitoes. Only basic computer skills are needed. After the initial expenditure for
an NIR spectrometer ($40,000 USD), the technique becomes more cost-effective than ovary dissection after about 40,000 samples have been analyzed. During collection of the egg-laying status of wild mosquitoes used in this thesis study, a technician was paid about $20 USD dollars per day of dissection (payments according to Ifakara Health Institute working manual). Each day, an average of 40 mosquitoes were dissected. Hence, the 927 mosquitoes used in this study required about 30 days and $480 USD. The NIRS scans used in this study required about one man-day and cost $40 USD. For comparison, NIR scanning of lab-reared mosquitoes was done at a rate of about 200 mosquitoes per hour.

5.2 Recommendations for Future Studies

If any doubts remain on the accuracy of NIRS, we recommend the following studies.

1. Semi-field experiments to imitate the wild environment. Mosquitoes to train and test the model would be raised in a screen house [17], which allows mosquitoes to grow in an environment close to the wild environment. During such an experiment, mosquitoes will not be given a sugar solution, but they will survive by feeding on juices from vegetation grown in the screen house.

2. Performing Polovodova ovary dissection [4] instead of Detinova dissection to address the concern we raised in Chapter Four whether mosquitoes who had
laid eggs were correctly classified by our model because Polovodova ovary
dissection tells the number of times a mosquito has laid eggs. From the
knowledge of egg-laying cycle, mosquitoes cannot have laid eggs twice when
they are less than seven days old. Therefore, if the model correctly estimates
egg-laid mosquitoes to be less than seven days, then these mosquitoes should
have laid eggs once.

3. We suggest marking a large number of lab-reared mosquitoes and releasing
them into the wild to survive in a natural environment following by
re-capturing them. Because they will be marked, once re-captured, their age
can be traced from the day they were released and the day they were captured.
This will provides labels for wild mosquitoes to train and validate the model.

4. Following available knowledge and tools to analyse protein expression patterns
associated with age, we suggest randomly selecting wild mosquitoes whose age
has been classified by our model and analyse their protein expression profile
associated with age.
BIBLIOGRAPHY


