MACHINE LEARNING AND DATA MINING-BASED METHODS TO ESTIMATE PARITY STATUS AND AGE OF WILD MOSQUITO VECTORS OF INFECTIOUS DISEASES FROM NEAR-INFRARED SPECTRA

by

Masabho Peter Milali, M.S.

A Dissertation Submitted to the Faculty of the Graduate School, Marquette University, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

May 2020
ABSTRACT

MACHINE LEARNING AND DATA MINING-BASED METHODS TO ESTIMATE PARITY STATUS AND AGE OF WILD MOSQUITO VECTORS OF INFECTIOUS DISEASES FROM NEAR-INFRARED SPECTRA

Masabho Peter Milali, M.S.
Marquette University, 2020

Previous studies show that a trained partial least square regresser (PLSR) from near-infrared spectra classify laboratory and semi-field raised mosquitoes into less than or ≥ to seven days old with an average accuracy of 80%. This dissertation demonstrates that training models on near-infrared spectra (NIRS) using artificial neural network (ANN) as an architecture yields models with higher accuracies than training models using partial least squares (PLS) as an architecture. In addition, irrespective of the model architecture used, direct training of a binary classifier scores higher accuracy than training a regresser and interpreting it as a binary classifier. Furthermore, for the first time, this dissertation shows that training ANN models on autoencoded near-infrared spectra yields models that estimate parity status of wild mosquitoes with an accuracy of ≈93%, which is strong enough to support NIRS models as an alternative to ovary dissections. Results from this dissertation also show that there is no significant difference between spectra collected from semi-field raised and wild mosquitoes of the same species, supporting the on-going practice of training models on semi-field raised mosquitoes to estimate the age class in days of wild mosquitoes. Finally, the study shows that an ANN model trained on semi-field mosquitoes classifies wild mosquitoes into either less than or ≥ to seven days old with an average accuracy of 76%. In conclusion, the results in this dissertation strongly suggest the use of ANNs as a suitable architecture to train models that estimate parity status and age in days of wild mosquito vectors of infectious diseases. The results further suggest near-infrared spectroscopy as an appropriate alternative tool to estimate different parameters of mosquito vectors of infectious diseases.
ACKNOWLEDGMENTS

Masabho Peter Milali, M.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, Dr. Naveen Bansal, and Dr. Serdar Bozdag for their support towards the successful completion of this dissertation.

Without the financial support from the Marquette Graduate School, the GasDay project, and Marquette Energy Analytics, which offered me scholarships for graduate studies, this work would not have been possible. Special thanks go to the Marquette MSCS department for accepting my application, which granted me 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-Lord and Dr. Fredros Okumu, who later accepted me into their project, which not only lead me into my dissertation topic, but also allowed me to use the data. I am grateful to them and to the Ifakara Institute.

I would also like to thank the GasDay lab lead by Dr. Richard Povinelli and Marquette Energy Analytics LLC lead by Thomas Quinn and Dr. Ronald Brown for putting their resources at my disposal during my PhD study - Thank you. To my colleagues and friends both in the GasDay lab and at Marquette Energy Analytics, 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, Daniella, and Joshua, we did it! Thank you for your understanding and for the support you have given me; I am glad to have you.
# TABLE OF CONTENTS

## ACKNOWLEDGMENTS

## LIST OF TABLES

## LIST OF FIGURES

### CHAPTER 1 Application of Machine Learning and Data Mining Techniques to Mosquito Near-infrared Spectra toward Eliminating Malaria Transmission

1.1 Statement of the problem .................................. 1

1.2 Status of the problem ........................................ 2

1.3 Statement of the materials .................................. 3

1.4 Statement of the approach .................................. 4

1.4.1 To explore ways to improve the current accuracy of PLS models for estimating the age of mosquitoes (Chapter 2) .... 7

1.4.2 To train models that estimate the parity status of wild mosquitoes (Chapter 3) ........................................ 8

1.4.3 To determine whether NIR spectra collected from semi-field raised mosquitoes differ from spectra collected from wild mosquitoes of the same species (Chapter 4) . ......................... 9

1.4.4 To train NIRS models that estimate the age in days of wild mosquitoes (Chapter 5) ..................................... 11

1.4.5 Conclusions and recommendations for future studies (Chapter 6) 13

### CHAPTER 2 Age Grading *An. gambiae s.s* and *An. arabiensis* Using Near Infrared Spectra and Artificial Neural Networks

2.1 Introduction to significance of knowing age of mosquitoes .......................... 17

2.2 Materials and methods ........................................ 21

2.2.1 Ethics approval ........................................... 21

2.2.2 Mosquito and spectra collection ........................... 21
<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.3 Model training</td>
<td>22</td>
</tr>
<tr>
<td>2.3 Results</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>33</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>47</td>
</tr>
<tr>
<td>2.6 Transition to Chapter 3</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to the problem of the study</td>
<td>52</td>
</tr>
<tr>
<td>Ethics approvals</td>
<td>54</td>
</tr>
<tr>
<td>Data</td>
<td>54</td>
</tr>
<tr>
<td>Model training and testing</td>
<td>56</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>63</td>
</tr>
<tr>
<td>Conclusion</td>
<td>73</td>
</tr>
<tr>
<td>Transition to Chapter 4</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to mosquito age estimation</td>
<td>76</td>
</tr>
<tr>
<td>Overview of cluster analysis</td>
<td>79</td>
</tr>
<tr>
<td>Data collection and processing</td>
<td>82</td>
</tr>
<tr>
<td>Cluster analysis and results</td>
<td>84</td>
</tr>
<tr>
<td>Discussion</td>
<td>99</td>
</tr>
<tr>
<td>Conclusion - No difference between spectra from semi-field and wild mosquitoes</td>
<td>102</td>
</tr>
<tr>
<td>Transition to Chapter 5</td>
<td>103</td>
</tr>
</tbody>
</table>

**CHAPTER 3** An Autoencoder and Artificial Neural Network-based Method to Estimate Parity Status of Wild Mosquitoes from Near-infrared Spectra

**CHAPTER 4** Do NIR Spectra Collected from Laboratory-reared Mosquitoes Differ from Those Collected from Wild Mosquitoes?
CHAPTER 5  An Artificial Neural Network Model to Estimate Age Class in Days of Wild *An. arabiensis*  

5.1 Introduction to mosquito age estimates from ovary dissection . . . . . . 106  
5.2 Material and method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108  
5.3 Results and discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111  
5.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120  

CHAPTER 6  Conclusions and Future Work  

6.1 General conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122  
6.2 Future work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125  

6.2.1 Training ANN models on autoencoded NIR spectra to estimate parameters of wild mosquitoes other than parity status . . . . . 125  
6.2.2 Applying transfer learning techniques to train a deep neural network model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126  
6.2.3 Developing a graphical user interface (GUI) . . . . . . . . . . . . . 128  
6.3 Concluding statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 129  

Bibliography  

BIBLIOGRAPHY
### LIST OF TABLES

1.1 List and summary of mosquito datasets used to test the reproducibility of results from our models. Numbers in brackets are references of the studies where the datasets are originally published.  

1.2 Number of mosquitoes per age group in each dataset used to test the reproducibility of our study.  

2.1 Performance analysis of PLS and ANN regression models on estimating age of *An. gambiae* and *An. arabiensis*. Results from ten-fold Monte Carlo cross-validation.  

2.2 Mean actual age estimation of mosquitoes in out-of-sample test sets by ANN and PLS regression models. Column ‘N’ represents the number of mosquitoes in each age group.  

2.3 Percentage of mosquitoes in each age group correctly classified when ANN and PLS regression models were interpreted as binary classifiers to classify mosquitoes into either < 7 days old or ≥ 7 days old.  

2.4 Reproducibility analysis of PLS and ANN regression models for estimating the age of *An. gambiae* and *An. arabiensis* in different datasets already used in other publications. DS1 - DS8 are mosquito datasets described in Tables 1.1 and 1.2. Results are from ten-fold Monte Carlo cross-validation.  

2.5 Performance analysis of PLS and ANN regressors for estimating the age of *Aedes albopictus*, *Wolbachia*-free, and *Wolbachia*-infected male and female *Aedes aegypti*. DS9 - DS17 are mosquito datasets described in Tables 1.1 and 1.2. Results are from ten-fold Monte Carlo cross-validation.  

2.6 Comparison of the accuracy of ANN and PLS classification models on ten replicates.  

2.7 Comparison of accuracies between directly trained binary classifiers and regressors interpreted as binary classifiers. Results from ten-fold Monte Carlo cross-validation. DTBC: Directly trained binary classifier; RIBC: Regresser interpreted as binary classifier.
2.8 Comparison of the accuracy of ANN and PLS classification models on An. gambiæ and An. arabiensis in datasets from other published studies described in Table 1.1.

2.9 Comparison of the accuracies of directly trained ANN and PLS classification models on Aedes aegypti and Aedes albopictus in datasets from other published studies described in Table 1.1.

2.10 Results when both regression and directly trained binary classifiers trained on IFA-GA and IFA-ARA datasets were tested on DS1 - 4 and DS7 - 8 (described in Table 2.1), respectively, as independent test sets.

3.1 Accuracies of reconstructing original feature spaces from encoded feature spaces. MSE = mean square error.

3.2 Performance of an ANN model trained on 75% of mosquito spectra with 1851 features (before autoencoder) and tested on the remaining 25% spectra (out of the sample testing).

3.3 Confusion matrices showing accuracies of the models in absolute values when the models were trained on spectra before feature reduction by autoencoder. A) Minepa-ARA, B) Muleba-GA, C) Burkina-GA and D) Muleba-Burkina-GA. Results from the last Monte-Carlo cross validation.

3.4 Performance of an ANN model trained on 75% of the encoded mosquito spectra (10 features) and tested on the remaining 25% of the encoded mosquito spectra.

3.5 Confusion matrices showing accuracies of the models in absolute values when the models were trained on spectra after feature reduction by autoencoder. A) Minepa-ARA, B) Muleba-GA, C) Burkina-GA and D) Muleba-Burkina-GA. Results from the last Monte-Carlo cross validation.

3.6 Independent testing of ANN models trained on the Muleba-GA and on the Burkina-GA encoded datasets.

4.1 Interpretation of the SC for partitioning methods [138]
4.2 Number and type of mosquitoes in clusters when $k$-means and hierarchical clustering were applied to spectra with: Age of mosquitoes not controlled ($A_k$ and $A_h$, respectively); Age structure of semi-field reared mosquitoes controlled to match the published age structure of wild mosquitoes ($B_k$ and $B_h$, respectively) and; semi-field reared mosquitoes at age 3, 5, and 25-day old not included in the analysis ($C_k$ and $C_h$, respectively). $X^2 = \text{computed chi-square}$ 88

4.3 Number of mosquitoes in clusters when 80 spectra collected from wild mosquitoes were randomly selected and maintained for the rest of the analysis, while changing the age of the semi-field reared mosquitoes. Av.SC: Average silhouette coefficient; $X^2$: Chi square 93

4.4 Number and type of mosquitoes in leaf nodes of the hierarchical tree: A) Age of mosquitoes not controlled; B) Age of mosquitoes controlled by selecting age of laboratory-reared mosquitoes to fit the published age distribution of wild mosquitoes; C) Laboratory-reared mosquitoes at age 3, 5, and 25-day old excluded in the analysis. 98

5.1 Performance of ANN binary classifiers on estimating the age class in days of 25% out-of-sample semi-field raised *An. arabiensis*. TP: Number of semi-field *An. arabiensis* correctly classified as $\geq 7$ days old; TN: Number of semi-field *An. arabiensis* correctly classified as $< 7$ days old; P: Total number of semi-field *An. arabiensis* that are $\geq 7$ days old; N: Total number of semi-field *An. arabiensis* that are $< 7$ days old. 112

5.2 Performance of the ANN binary classifier (model-seven) trained on semi-field raised *An. arabiensis* ($N = 653$) on classifying wild *An. arabiensis* ($N = 912$) into less than or $\geq 7$ days old. 114

5.3 The number of parous and nulliparous wild *An. arabiensis* in each age class estimated by model-seven after the model was applied to classify wild *An. arabiensis* into either less than or $\geq 7$ days old. 114

5.4 Accuracy of model-five in estimating age class of 25% out-of-sample semi-field raised *An. arabiensis*. TP: Number of semi-field *An. arabiensis* correctly classified as $\geq 5$ days old; TN: Number of semi-field *An. arabiensis* correctly classified as $< 5$ days old; P: Total number of semi-field *An. arabiensis* that are $\geq 5$ days old; N: Total number of semi-field *An. arabiensis* that are $< 5$ days old. 116
5.5 The number of parous and nulliparous wild *An. arabiensis* in each age class estimated by model-five after the model was applied to classify wild *An. arabiensis* into either less than or $\geq 5$ days old.
LIST OF FIGURES

2.1 The percentage of variance explained in the dependent variable against the number of PLS components: A) *An. gambiae*  
B) *An. arabiensis* ........................................... 23

2.2 Reproducing our analysis on different datasets (black DS = *An. gambiae*; green DS = *An. arabiensis*; red DS = *Ae. aegypti* and *Ae. albopictus*) ................................................................. 27

2.3 ANN and PLS models trained on IFA-ARA and IFA-GA datasets were tested on independent datasets. ............. 28

2.4 PLS (A and C) and ANN (B and D) regression models, estimating actual age of training and testing samples of *An. gambiae* (A and B) and *An. arabiensis* (C and D), respectively 30

2.5 Regression coefficients weights against wavelengths: A) *An. gambiae* B) *An. arabiensis* ................................. 31

2.6 Box plots when PLS (A and C) and ANN (B and D) were applied to estimate actual age of out of the sample *An. gambiae* (A and B) and *An. arabiensis* (C and D), respectively. .... 32

2.7 Number of *An. gambiae s.s* (A and B) and *An. arabiensis* (C and D) in two age classes (less than or greater/equal seven days) when PLS (A and C) and ANN (B and D) regression models, respectively, interpreted as binary classifiers. ....... 33

2.8 The consistency in accuracies of directly trained PLS (A and C) and ANN (B and D) binary classifiers for estimating age classes of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in both training and testing sets ........................................ 38

2.9 Box plot of directly trained PLS (A and C) and ANN (B and D) binary classifiers for estimating age classes of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in out-of-sample testing sets. .......................... 39
2.10 The number of correct and false predictions in each estimated age class when directly trained PLS (A and C) and ANN (B and D) binary classifiers were applied to classify the ages of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in testing sets. Results from ten replicates. ........................................ 40

2.11 Error distribution per actual age of *An. gambiae* and *An. arabiensis* when ANN and PLS regressors applied to estimate the actual ages of mosquitoes in training and test data sets, showing a uniform distribution of errors (un-biased estimating) across actual ages of mosquitoes for the ANN regressor and an un-uniform distribution of errors (biased estimating) for the PLS regressor ................................................................. 47

2.12 Error distribution per actual age of *An. gambiae* and *An. arabiensis* when ANN and PLS regressors applied to estimate the actual ages of mosquitoes in training and test data sets, showing a uniform distribution of errors (un-biased estimating) across actual ages of mosquitoes for the ANN regressor and an un-uniform distribution of errors (biased estimating) for the PLS regressor ................................................................. 48

3.1 Training and testing ANN model on spectra preprocessed according to Mayagaya et al. [75]. “M” is either Minepa-ARA, Muleba-GA, Burkina-GA, or Muleba-Burkina-GA .................................................. 57

3.2 Autoencoder reducing feature space dimension .......................... 59

3.3 ANN model trained on a dataset with an encoded feature space 60

3.4 Reducing spectra feature space using an autoencoder and re-constructing original feature spaces from their respective encoded feature spaces (reconstruction accuracies presented in Table 3.1). Figures generated from MATLAB ........................................ 61

3.5 Training and testing of ANN model on autoencoded spectra. M is either Minepa-ARA, Muleba-GA, Burkina-GA, or Muleba-Burkina-GA dataset ................................................................. 62

3.7 Box plots showing results when ANN models trained on 75% of spectra before the autoencoder was applied and tested on the remaining spectra (25%) (out-of-sample testing). A, B, C, and D represent results for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba & Burkina-GA (mosquitoes in the Muleba-GA and the Burkina-GA datasets combined) datasets, respectively.

3.8 ROC curves (AUCs presented in the last row of Table 3.2) showing results when ANN models trained on 75% of spectra before the autoencoder was applied and tested on the remaining spectra (25%) (out-of-sample testing). A, B, C, and D represent results for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba & Burkina-GA (mosquitoes in Muleba-GA and Burkina-GA datasets combined) datasets, respectively.

3.9 Box plots showing results when ANN models trained on 75% of encoded spectra in datasets were tested on the remaining encoded spectra (25%). A, B, C, and D represent results for Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA, and Encoded-Muleba & Burkina-GA (mosquitoes in Encoded-Muleba-GA and Encoded-Burkina-GA datasets combined) datasets, respectively.

3.10 ROC curves (AUCs presented in the last row of Table 2) showing results when ANN models trained on 75% of encoded spectra were tested on the remaining encoded spectra (25%). A, B, C, and D represent results for Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA and Encoded-Muleba & Burkina-GA (mosquitoes in Encoded-Muleba-GA and Encoded-Burkina-GA datasets combined) datasets, respectively.

4.1 Two-dimensional plot of clusters using absorbances at 500 nm and 501 nm, when the age of mosquitoes was not controlled

4.2 Box plots of silhouette coefficients and bar graphs of percentage of mosquitoes, respectively, showing the quality and distribution of laboratory-reared and wild mosquitoes in clusters after k-means analysis. Panels A and B, age of mosquitoes was not controlled ($p = 0.01$); C and D, age structure of laboratory-reared mosquitoes was controlled to match the published age structure of wild mosquitoes ($p = 0.57$); E and F, laboratory-reared mosquitoes at 3, 5, and 25-day old were not included in the analysis ($p = 0.26$). $P$ stands for $p$-value, and $N$ for the number of mosquitoes.

4.3 Number of semi-field reared mosquitoes per age, selected to imitate the published age distribution [4, 11, 17, 70, 32, 142] of wild mosquitoes
4.4 Illustration of the second method used to control number of mosquitoes per age during \( k \)-means approach two clustering ...................... 92

4.5 Two-dimensional plot of clusters using the first and second PLS components (Panel A), and box plots, showing the silhouette coefficient of each spectrum (object) in its associated cluster (Panel B) when partial least squares was applied to reduce the data dimension before clustering ...................... 96

4.6 Hierarchical tree and bar graphs showing distributions of semi-field reared and wild mosquitoes in clusters formed by hierarchical cluster analysis: In panels A and B, the age of mosquitoes was not controlled \( (p < 0.01) \); in C and D, the age structure of semi-field reared mosquitoes was fit to an exponential decay distribution to match the published age structure of wild mosquitoes \( (p = 0.76) \); and in E and F, semi-field reared mosquitoes at 3, 5, and 25-days old were omitted from the analysis \( (p = 0.13) \). ..................................... 97

5.1 Percentage of nulliparous and parous mosquitoes in each age class estimated by model-seven after it was applied to classify wild \textit{An. arabiensis} (\( N = 912 \)) into either less than or \( \geq 7 \) days old .................................................. 115

5.2 Receiver operating curve (ROC) with area under the curve (AUC) presenting the performance of model-five when was applied to classify 25\% out-of-sample semi-field raised \textit{An. arabiensis} into less than or \( \geq 5 \) days old ................. 117

5.3 Percentage of nulliparous and parous mosquitoes in each age class estimated by model-five after it was applied to classify wild \textit{An. arabiensis} (\( N = 912 \)) into either less than or \( \geq 5 \) days old .................................................. 118

5.4 Using the Jaccard similarity coefficient to determine similarity in outputs from models trained on semi-field raised mosquitoes to estimate the age class of wild mosquitoes and from Detinova ovary dissection ........................................ 120
This chapter provides a general overview of the dissertation.

1.1 Statement of the problem

Malaria is a vector-borne parasitic disease transmitted by mosquitoes of the genus *Anopheles*. Approximately 228 million malaria cases and 405,000 deaths caused by malaria were recorded globally in 2018 [144]. Malaria transmission occurs when mosquitoes acquire *Plasmodium* parasites in an infected blood meal, survive long enough to support parasite development to the infectious sporozoite stage, and transmit it to a susceptible host [9]. Due to the long incubation period of 7-14 days required by the parasite to develop within the mosquito, malaria parasites are transmitted only by mosquitoes that are at least seven days old [9]. This makes mosquito survival and age critical factors to parasite transmission.

Vector management is one of the interventions in the fight against malaria transmission. It mainly includes the use of long-lasting insecticide treated nets (LLINs) and indoor residual spraying (IRS) with insecticides [72, 111]. Control and evaluation of LLINs and IRS involves determination of the age structure of
mosquito populations in areas where these interventions are applied [21]. If the interventions are working, the expectation is to see a mosquito population with the ratio of old mosquitoes to young mosquitoes decreasing.

The current techniques to estimate mosquito age are based on dissection of their ovaries to determine whether they have laid eggs. Those found to have laid eggs are assumed to be older than those found to not have laid eggs [29]. However, the difficulty and laborious dissections involved with this technique limits its application to only a few experts who end up working with small numbers of mosquitoes. In addition, this method cannot estimate age chronologically when needed. It is limited to age classification. As a result, new approaches which can complement ovary dissections are needed.

1.2 Status of the problem

One promising approach to complement ovary dissection is near infrared spectroscopy (NIRS). NIRS is a high throughput technique, measuring the energy absorbed by biological samples [7, 15, 131]. NIRS has been demonstrated to estimate age, species, and infectious state of laboratory reared and semi-field raised mosquitoes [35, 62, 73, 75, 76, 81, 95, 120, 121, 127, 128] with an average accuracy of 85%. This accuracy offers room for improvement. Also, despite the demonstrated ability of NIRS to estimate the age in days of laboratory and semi-field raised
mosquitoes, it has not been tested on its ability to estimate the egg-laying status of mosquitoes and not comprehensively tested on wild mosquitoes, limiting the application of NIRS technology as an alternative tool to estimate parameters (such as age, species, parity status, and infectious state) of mosquito vectors of infectious diseases.

1.3 Statement of the materials

The materials used in this dissertation were near-infrared spectra scanned from: i) laboratory reared *An. gambiae s.s* and semi-field raised *An. arabiensis* collected from an insectary and semi-field systems owned and maintained by the Ifakara Health Institute (IHI) in Tanzania; ii) wild *An. arabiensis* and *An. gambiae s.s* collected in households located in villages found in southeastern and northwestern Tanzania. Scanning of mosquitoes was performed at IHI using a near-infrared spectroscopy (NIRS) machine on loan from the United States Department of Agriculture (USDA), facilitated by Dr. Floyd Dowell, a research scientist at USDA. We also used published and publicly available datasets summarized in Tables 1.1 and 1.2.
Table 1.1: **List and summary of mosquito datasets used to test the reproducibility of results from our models.** Numbers in brackets are references of the studies where the datasets are originally published.

<table>
<thead>
<tr>
<th>Species</th>
<th>Dataset ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An. gambiae</em> [62]</td>
<td>DS1</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; generation mosquitoes emerged from wild larvae collected in 2013 at Soumouso (DS1) and in 2014 at Kodeni (DS2) in southwestern Burkina Faso and reared under ambient conditions</td>
</tr>
<tr>
<td></td>
<td>DS2</td>
<td>Mosquitoes from a colony established in 2015 with original larvae collected in Burkina Faso (DS3) and from a colony established in 1975 (DS4). Both colonies reared at Colorado State University (CSU) at 28±2°C and 80% humidity under a 14:10 light:dark photoperiod</td>
</tr>
<tr>
<td></td>
<td>DS3</td>
<td>DS4</td>
</tr>
<tr>
<td><em>An. arabiensis</em> [95, 121]</td>
<td>DS7 [95]</td>
<td>Reared at Ifakara Health Institute in semi-field systems under ambient conditions. Spectra collection at 3, 5, 8, and 11 days old using QualitySpec Pro Spectrometer (ASD Inc, Boulder, CO). Killed using chloroform before spectra collection</td>
</tr>
<tr>
<td></td>
<td>DS8 [121]</td>
<td>Wild larvae and pupae collected in Zanzibar from different mosquito breeding sites and reared under ambient conditions. Spectra collection at 1, 3, 5, 7, 9, and 14 days old using LabSpec 5000 NIR spectrometer (ASD Inc, Boulder, CO). RNAlater to preserve samples before spectra collection. Pyrethroid resistant.</td>
</tr>
<tr>
<td><em>Aedes aegypti</em> [128]</td>
<td>DS9</td>
<td>Reared at the insectary of QIMR Berghofer Medical Research Institute, Australia at 27°C, 70% humidity, 12:12 hr day:night lighting</td>
</tr>
<tr>
<td></td>
<td>DS10</td>
<td>Wolbachia-free females</td>
</tr>
<tr>
<td></td>
<td>DS11</td>
<td>wMelPop infected males</td>
</tr>
<tr>
<td></td>
<td>DS12</td>
<td>wMelPop infected female</td>
</tr>
<tr>
<td></td>
<td>DS13</td>
<td>wMel infected males</td>
</tr>
<tr>
<td></td>
<td>DS14</td>
<td>wMel infected female</td>
</tr>
<tr>
<td></td>
<td>DS15</td>
<td>DS9, DS11 and DS13 combined</td>
</tr>
<tr>
<td></td>
<td>DS16</td>
<td>DS10, DS12 and DS14 combined</td>
</tr>
<tr>
<td><em>Aedes albopictus</em> [126]</td>
<td>DS17</td>
<td>Reared at the insectary in QIMR Berghofer Medical Research Institute, Australia at 27°C, 70% humidity with 12:12 hr day:night lighting and 30 min dawn/dusk periods. Spectra collection at 3, 7, 9, 13, 16, 20, and 25 days old using LabSpec 4Si NIR spectrometer model (ASD Inc, Boulder, CO). Preserved in RNA-later before spectra collection</td>
</tr>
</tbody>
</table>

### 1.4 Statement of the approach

In this dissertation, four objectives were accomplished. For each of the objectives, a manuscript was generated for publication in a peer-reviewed journal. Below are the
Table 1.2: Number of mosquitoes per age group in each dataset used to test the reproducibility of our study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1d</th>
<th>3d</th>
<th>5d</th>
<th>6d</th>
<th>7d</th>
<th>8d</th>
<th>9d</th>
<th>10d</th>
<th>11d</th>
<th>12d</th>
<th>13d</th>
<th>14d</th>
<th>15d</th>
<th>16d</th>
<th>17d</th>
<th>19d</th>
<th>20d</th>
<th>21d</th>
<th>25d</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>-</td>
<td>48</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>48</td>
<td>-</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS2</td>
<td>-</td>
<td>39</td>
<td>-</td>
<td>40</td>
<td>-</td>
<td>-</td>
<td>37</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>-</td>
<td>-</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS3</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>-</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS4</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS5</td>
<td>-</td>
<td>87</td>
<td>-</td>
<td>90</td>
<td>-</td>
<td>-</td>
<td>85</td>
<td>-</td>
<td>-</td>
<td>89</td>
<td>-</td>
<td>-</td>
<td>66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS6</td>
<td>-</td>
<td>137</td>
<td>-</td>
<td>139</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>106</td>
<td>-</td>
<td>-</td>
<td>108</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS7</td>
<td>-</td>
<td>104</td>
<td>-</td>
<td>122</td>
<td>-</td>
<td>-</td>
<td>92</td>
<td>-</td>
<td>-</td>
<td>209</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS8</td>
<td>9</td>
<td>61</td>
<td>51</td>
<td>101</td>
<td>47</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS9</td>
<td>25</td>
<td>30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>76</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>95</td>
<td>-</td>
<td>-</td>
<td>55</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>DS10</td>
<td>72</td>
<td>110</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>71</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>114</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>DS11</td>
<td>43</td>
<td>48</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46</td>
<td>-</td>
<td>-</td>
<td>51</td>
</tr>
<tr>
<td>DS12</td>
<td>40</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>53</td>
</tr>
<tr>
<td>DS13</td>
<td>48</td>
<td>60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>29</td>
</tr>
<tr>
<td>DS14</td>
<td>52</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>-</td>
<td>44</td>
</tr>
<tr>
<td>DS15</td>
<td>116</td>
<td>138</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>76</td>
<td>88</td>
<td>-</td>
<td>-</td>
<td>66</td>
<td>-</td>
<td>-</td>
<td>96</td>
<td>-</td>
<td>95</td>
<td>80</td>
<td>47</td>
<td>55</td>
<td>36</td>
</tr>
<tr>
<td>DS16</td>
<td>164</td>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>71</td>
<td>91</td>
<td>-</td>
<td>-</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>94</td>
<td>-</td>
<td>114</td>
<td>97</td>
<td>49</td>
<td>30</td>
<td>73</td>
</tr>
<tr>
<td>DS17</td>
<td>49</td>
<td>51</td>
<td>51</td>
<td>53</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>102</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>48</td>
<td>-</td>
<td>-</td>
<td>54</td>
<td>-</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>

objectives and the titles of the manuscripts generated from each of the objectives and their status:

1. To explore ways to improve the current accuracy of partial least squares regression (PLSR) models for estimating the age of mosquitoes - Paper title: Age grading *An. gambiae* s.s and *An. arabiensis* using near-infrared spectra and artificial neural networks (published as [83])

2. To train models that estimate parity status of wild mosquitoes - Paper title: An autoencoder and artificial neural network-based method to estimate parity status of wild mosquitoes using near-infrared spectra (Published in bioRxiv and under review in PLOS ONE)
3. To determine whether NIR spectra collected from semi-field raised mosquitoes differ from spectra collected from wild mosquitoes of the same species - Paper title: Do NIR spectra collected from laboratory-reared mosquitoes differ from those collected from wild mosquitoes? (published as [84])

4. To train models on spectra from semi-field raised mosquitoes to estimate the age in days of wild mosquitoes - Paper title: An artificial neural network model applied to estimate age class in days of wild An. arabiensis (In preparation for submission to Malaria Journal)

With an exception of the last chapter (Chapter Six), the rest of the chapters in this dissertation report are slightly edited manuscripts (either published, submitted, or ready to be submitted), each with a stand-alone introduction, materials, methods, results, discussion, conclusion, and recommendations sections. The last chapter (Chapter Six) provides general conclusions from all four objectives and recommendations for future studies and applications of the near-infrared technology.

Next, in this chapter, we summarize our contributions represented by the findings from each of the four objectives.
1.4.1 To explore ways to improve the current accuracy of PLS models for estimating the age of mosquitoes (Chapter 2)

**Contribution:** In Chapter Two, we offer ways to improve on the state-of-the-art accuracy of NIRS models trained to estimate age of mosquito vectors of infectious diseases.

Previous studies [35, 75, 76, 119, 120, 121] trained regression models on NIR spectra using partial least squares and interpreted them as binary classifiers to classify laboratory reared and semi-field raised mosquitoes into either less than or greater than or equal to seven days old with an average accuracy of 80%. Since the output from the regression model is age as a continuous variable, the process of interpreting this model output into a binary output can introduce errors affecting the accuracy of the model. Direct training of a binary classifier instead of training a regresser and interpreting it as a binary classifier may improve the current reported accuracy of the NIRS models. Therefore in Chapter Two, using partial least squares as a model architecture, we directly trained a binary classifier and compared the accuracy with the trained regresser translated as a binary classifier. In addition, the model architecture often contributes to the performance of the model [91]. Studies [12, 58, 88, 150] show that models trained using artificial neural networks as a model architecture perform better than models trained using partial least squares as a model architecture. Hence, we further trained regressers and binary classifiers
using both artificial neural networks and partial least squares as model architectures and compared their performances.

We find training both binary classifiers and regressers using ANN architectures yields models with higher accuracies than when similar models are trained using partial least squares. Furthermore, irrespective of the model architecture, direct training of the binary classifiers yields a model with higher accuracy than training a regresser and interpreting it as a binary classifier. Hence, we recommend using ANN architectures to train models that estimate the age of mosquitoes. In case an age class is desired, then direct training of a binary classifier is recommended over training a regresser and interpreting it as a binary classifier.

In the next Chapter, we train models that estimate parity status of wild mosquitoes.

1.4.2 To train models that estimate the parity status of wild mosquitoes (Chapter 3)

Contribution: In Chapter Three, for the first time in the literature, we present models that estimate the parity status (egg laying status) of wild mosquito vectors of infectious diseases.

The parity status of mosquitoes is important in determining the capacity of a population of wild mosquitoes to transmit diseases. A population of wild
mosquitoes with more parous (laid eggs) mosquitoes than nulliparous (not laid eggs) mosquitoes is more potentially infectious than a population with more nulliparous mosquitoes than parous mosquitoes. Parity status also is used to control and evaluate interventions to control mosquito vectors of diseases such as insecticide treated nets (ITNs) and indoor residual spraying (IRS). While in areas with fewer parous mosquitoes, the interventions in place probably are working; in areas with more parous mosquitoes, the interventions are not effective. The current method to score the parity status of wild mosquitoes involves the dissection of mosquito ovaries to determine whether they have laid eggs. The dissection process is tedious, time consuming, and requires skilled personnel. We train models on near-infrared spectra using autoencoders and an artificial neural network to estimate the parity status of wild mosquitoes. The models scored an average accuracy of 93%, suggesting an alternative method to estimate the parity status of wild mosquitoes.

Next in Chapter Four, we determine if there is a significant difference between NIR spectra collected from semi-field reared and wild collected mosquitoes of the same species.

1.4.3 To determine whether NIR spectra collected from semi-field raised mosquitoes differ from spectra collected from wild mosquitoes of the same species (Chapter 4)

**Contribution:** In Chapter Four, for the first time in the literature, we offer
evidences that supports the on-going practice of training models on NIR spectra collected from semi-field reared mosquitoes to estimate different parameters of wild mosquito vectors of infectious diseases.

The success of machine learning NIR spectra techniques applied to estimate the age in days of laboratory and semi-field raised mosquitoes is not meaningful if we cannot apply NIRS models to estimate the age in days of wild mosquitoes. Neither laboratory reared nor semi-field raised mosquitoes transmit diseases; wild mosquitoes do. Lacking samples of wild mosquitoes whose ages in days are known to train and test the model limits development and scoring of the NIRS models that estimate the age in days of wild mosquitoes. Methods such as mark-release-recapture \cite{26, 92} that can offer samples of wild mosquitoes with known age in days are very expensive and time consuming. Alternatively, models developed from spectra collected from semi-field raised mosquitoes currently are applied to estimate the age in days of wild mosquitoes \cite{62, 81, 128}. This practice is appropriate only if there is no significant difference between spectra collected from semi-field raised and wild mosquitoes of the same species. No study has been done to support this generalization. Wild mosquitoes are exposed to different food materials and light intensities compared to semi-field raised mosquitoes. This may affect the characteristics of their exoskeletons, especially hardness. The hardness of the exoskeleton influences the amount of near infrared light absorbed by a mosquito.
As a result, it might be that two mosquitoes of the same age and species, but one from the semi-field system and another from the wild, yield significantly different NIR spectra.

We apply $k$-means and hierarchical clustering techniques to determine whether there is any significant difference between spectra collected from semi-field raised and wild collected *Anopheles arabiensis*. With the results from both $k$-means and hierarchical clustering, we fail to reject the null hypothesis that there is no significant difference between spectra collected from semi-field reared and wild collected *An. arabiensis* ($p = 0.25$ and $p = 0.13$ for $k$-means and hierarchical clustering, respectively). Hence, strengthening and supporting the ongoing practice [62, 81, 128] of training models on semi-field raised mosquitoes to estimate the age in days of wild mosquitoes.

Next in Chapter Five, we further train models on NIR spectra collected from semi-field reared *An. arabiensis* to estimate age in days of wild collected *An. arabiensis*.

### 1.4.4 To train NIRS models that estimate the age in days of wild mosquitoes (Chapter 5)

**Contribution:** In this Chapter, for the first time in the literature, we show the accuracy of ANN models trained on semi-field *An. arabiensis* to estimate age class in days of wild collected *An. arabiensis*. 
Knowing from Chapter Four and in [84] that there is no significant difference between NIR spectra collected from semi-field raised and wild mosquitoes of the same species, in Chapter Five, we train models on semi-field raised mosquitoes to estimate the age in days of wild mosquitoes. Previous studies trained PLS models on semi-field raised An. gambiae and Aedes aegypti to estimate age class in days of wild An. gambiae and Aedes aegypti, respectively. In this Chapter (Chapter Five), we train an ANN model on semi-field An. arabiensis and apply it to wild An. arabiensis. Furthermore, since in Chapter Two and in [83], we show that ANN models perform better than PLS models, in Chapter Five, unlike studies in [62, 81, 128], we train ANN models instead of PLS models. Lacking age in days labels for wild An. arabiensis, we scored our model using their parity status (egg laying status) labels with an assumption that wild An. arabiensis that have laid eggs should be classified by our model (trained on semi-field An. arabiensis) as greater or equal to seven days. This is because, according to the mosquito reproduction cycle, the chances of a female mosquito having laid eggs when she is less than seven days old are very small [74, 89]. In addition we use the Jaccard similarity coefficient [54, 94] to score the accuracy of our model. Our model trained on semi-field raised An. arabiensis classified wild An. arabiensis into either less than or greater or equal to seven days with 74.7% accuracy. Also, according to the Jaccard similarity coefficient, the chances that our ANN model classifies wild An. arabiensis into the same class as ovary dissection is 75%. Since NIRS is a high-throughput
technique compared to ovary dissection, this accuracy and similarity is sufficient for practical use, especially when the factor of sample size is considered. NIRS with its high-throughput characteristic offers larger sample size that is more statistically acceptable to draw conclusions than sample sizes from ovary dissections.

Hence in Chapter Five, we recommend complementing ovary dissection with ANN models trained on semi-field raised mosquitoes to estimate age class in days of wild mosquito vectors of infectious diseases. Applying both ANN models and ovary dissections will provide more reliable age class estimates.

### 1.4.5 Conclusions and recommendations for future studies (Chapter 6)

This dissertation generally concludes that neural networks models estimate parity status and age class in days of wild mosquito vectors of infectious diseases with accuracies sufficient to either replace or complement the current technique to estimate both parity status and age class in days of wild mosquitoes. The study further concludes that the on-going practice of training models on NIR spectra collected from semi-field raised mosquitoes to estimate age class in days of wild collected mosquitoes is reasonable and appropriate. Lastly, in cases where age class is needed, this dissertation recommends direct training of a binary classifier over training a regressor and interpret it as a binary classifier. Direct trained binary classifiers score higher accuracies than regressors interpreted as binary classifiers.
We recommend future studies to explore ways to train models that can easily be modified and extrapolated to estimate different parameters of mosquitoes using NIR spectra with different characteristics than NIR spectra used to train and test the models. The current approach to train models yields models that perform only one specific objective (i.e., estimating either age or species but not both). In cases where other estimates such as parity status are required, new and different models are trained from scratch. The approach also yields models that perform poorly when extrapolated to estimate parameters of mosquitoes with characteristics (i.e., geographical regions; killing methods and version of the instrument used for spectra collection) other than the characteristics of the mosquitoes used to train the model. These limitations on the current ways to train models limit the application and scaling up of near-infrared technology as a tool to estimate parameters of mosquito vectors of infectious diseases.
CHAPTER 2

Age Grading An. gambiae s.s and An. arabiensis Using Near Infrared Spectra and Artificial Neural Networks

This chapter is adapted from [83], as published in *PLoS One*

Abstract

**Background:** Near infrared spectroscopy (NIRS) complements techniques to age-grade mosquitoes. NIRS classifies lab-reared and semi-field raised mosquitoes into less than seven days or greater than or equal to seven days old with an average accuracy of 80%, achieved by training a regression model using partial least squares (PLS) and interpreted as a binary classifier.

**Methods and findings:** We explored whether using an artificial neural network (ANN) analysis instead of PLS regression improves the accuracy of NIRS models for age-grading malaria-transmitting mosquitoes. We also explored if directly training a binary classifier instead of training a regression model and interpreting it as a binary classifier improves the accuracy. A total of 786 and 870 NIR spectra collected from laboratory-reared *An. gambiae* and *An. arabiensis*, respectively, were used and pre-processed according to previously published protocols. The ANN regression model scored root mean squared error (RMSE) of $1.6 \pm 0.2$ for *An.*
An. gambiae and 2.8 ± 0.2 for An. arabiensis; whereas the PLS regression model scored RMSE of 3.7 ± 0.2 for An. gambiae and 4.5 ± 0.1 for An. arabiensis. When we interpreted regression models as binary classifiers, the accuracy of the ANN regression model was 93.7 ± 1.0% for An. gambiae and 90.2 ± 1.7% for An. arabiensis; while the PLS regression model achieved an accuracy of 83.9 ± 2.3% for An. gambiae and 80.3 ± 2.1% for An. arabiensis.

It was also found that a directly trained binary classifier yields higher age estimation accuracy than a regression model interpreted as a binary classifier. A directly trained ANN binary classifier scored an accuracy of 99.4 ± 1.0% for An. gambiae and 99.0 ± 0.6% for An. arabiensis; while a directly trained PLS binary classifier scored 93.6 ± 1.2% for An. gambiae and 88.7 ± 1.1% for An. arabiensis.

We further tested the reproducibility of these results on independent mosquito datasets. ANN models scored higher estimation accuracies than PLS models. Regardless of the model architecture, directly trained binary classifiers scored higher accuracies than regression models translated as binary classifiers.

**Conclusion:** We recommend training models to estimate the age of An. arabiensis and An. gambiae using ANN model architectures (especially for datasets with at least 70 mosquitoes per age group) and direct training of binary classifier instead of training a regression model and interpreting it as a binary classifier.
2.1 Introduction to significance of knowing age of mosquitoes

Estimating the age of mosquitoes is one of the indicators used by entomologists for estimating vectorial capacity [38] and the effectiveness of an existing mosquito control intervention. Malaria is a vector-borne parasitic disease transmitted to people by mosquitoes of the genus *Anopheles*. The disease killed approximately 405,000 people in 2018 [144]. Mosquitoes contribute to malaria transmission by hosting and allowing the development to maturity of the malaria-causing *Plasmodium* parasite [9]. Depending on environmental temperature, *Plasmodium* takes 10-14 days in an *Anopheles* mosquito to develop fully enough to be transmitted to humans [9]. Therefore, knowing the age of a mosquito provides an indication of whether a mosquito is capable of transmitting malaria.

Knowing the age of a mosquito population is also important when evaluating the effectiveness of a mosquito control intervention. Commonly used vector control interventions such as insecticide treated nets (ITNs) and indoor residual spraying (IRS) reduce the abundance and the lifespan of a mosquito population to a level that does not support *Plasmodium* parasite development to maturity [72, 111]. Monitoring and evaluation of ITNs and IRS involves determining the 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 (ITNs or IRS)
intervention indicates that the intervention is working. On the other hand, if there are more old mosquitoes, the intervention is not working effectively.

The current techniques used to estimate mosquito age are based on a combination of ovary dissecting and conventional microscopy to determine their egg-laying history. Those found to have laid eggs are assumed to be older than those found to not have laid eggs [31]. This assumption can be misleading, as mosquitoes can be old but have not laid eggs and can be young (at least three days old) and have laid eggs. Dissection is laborious, difficult, and limited to only few experts. As a result, we need a new approach to address these limitations.

Different techniques such as a change in abundance of cuticular hydrocarbons [18, 52], transcriptional profiles [27, 79], and proteomics [122, 123] have been developed to age grade Anopheles mosquitoes. However, these techniques are still in early development stages and are limited to analyzing a small number of samples due to high analysis costs involved.

Near infrared spectroscopy (NIRS) is a complementary method to the current mosquito age grading techniques [75, 120]. NIRS is a high throughput 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 [36]; to age grade houseflies [101], stored-grain pests [102], and biting midges [110]; to differentiate between species and subspecies of termites [2]; to
estimate the age and to identify species of morphologically indistinguishable laboratory reared and semi-field raised *Anopheles gambiae* and *Anopheles arabiensis* mosquitoes [35, 75, 76, 119, 120, 121]; to estimate the age of *Aedes aegypti* mosquitoes [128]; and to detect and identify two strains of *Wolbachia pipientis* (wMelPop and wMel) in male and female laboratory-reared *Aedes aegypti* mosquitoes [127].

The current state-of-the-art of the accuracy of NIRS to classify the age of lab-reared *An. gambiae* and *An. arabiensis* is an average of 80% [35, 75, 76, 119, 120, 121]. This accuracy is based on a trained regression model using partial least squares (PLS) and interpreted as a binary classifier to classify mosquitoes into two age groups (less than seven days and greater or equal to seven days).

In this chapter, using a set of spectra collected from lab-reared and field collected *An. gambiae* and *An. arabiensis*, we explored ways to improve the reported accuracy of a PLS model for estimating age of mosquito vectors of infectious diseases. Selection of an architecture to train a model is one of the important factors influencing the accuracy of the model [91]. Studies [12, 58, 88, 150] compared the accuracies of artificial neural network (ANN) and PLS regression models concluding that ANN models generally perform better than PLS models. Therefore, using ANN [12, 50, 58] and PLS, we trained regression age models and compared results.
Since previous studies [35, 75, 76, 119, 120, 121] trained a regression model and interpreted it as a binary classifier (less than seven days and greater or equal to seven days), the interpretation process may introduce errors and compromise the accuracy of the model. We further trained ANN and PLS binary classifiers and compared their accuracies with the ANN and PLS regression models translated as binary classifiers.

The study finds that training of both regression and binary classification models using an artificial neural network architectures yields higher accuracies than when the corresponding models are trained using partial least squares model architectures. Also, regardless of the architecture of the model, training a binary classifier yields higher age class estimation accuracy than a regression model interpreted as a binary classifier.

We then tested the reproducibility of our results by applying similar analyses on different mosquito data sets from other published studies [62, 95, 121, 126, 128], whose data are freely available for reuse.
2.2 Materials and methods

This section first provides an ethical approval of the study. It further describes the obtaining and scanning of mosquitoes to collect NIR spectra. The section finally explains how models were trained.

2.2.1 Ethics approval

Permission for blood feeding laboratory-reared mosquitoes was obtained from the Ifakara Health Institute (IHI) Review Board, under Ethical clearance No. IHRDC/EC4/CL.N96/2004. Oral consent was obtained from each adult volunteer involved in the study. The volunteers were given the right to refuse to participate or to withdraw from the experiment at any time.

2.2.2 Mosquito and spectra collection

We used spectra of *Anopheles gambiae* (IFA-GA) mosquitoes collected at 1, 3, 5, 7, 9, 11, 15, and 20 days and *Anopheles arabiensis* (IFA-ARA) collected at 1, 3, 5, 7, 9, 11, 15, 20, and 25 days post emergence from the Ifakara Health Institute insectary. While *An. arabiensis* were reared in a semi-field system (SFS) at ambient conditions, *An. gambiae* were reared in a room made of bricks at controlled conditions. Adult mosquitoes were often provided with a human blood meal in a
week and 10% glucose solution daily. Mosquitoes were killed by freezing for 20
minutes and left to re-equilibrate to room temperature for approximately 30
minutes. Using a LabSpec 5000 NIR spectrometer with an integrated light source
(ASD Inc., Longmont, CO), we followed the protocol supplied by Mayagaya and
colleagues to collect spectra [75]. A total of 786 *An. gambiae* and 870 *An.
arabiensis* were scanned with at least 70 mosquitoes from each age group.

### 2.2.3 Model training

We first trained ANN and PLS regression models, scored and compared their
accuracies as regressors and then as binary classifiers. We further trained binary
classifiers and compared the accuracies with regressors interpreted as binary
classifiers. We used a two-tail t-test to test the hypothesis that there is a significant
difference in accuracies between ANN and PLS models, a one-tail t-test to test the
hypothesis that an ANN model scores higher accuracies than a PLS model.

In each species, we separately processed spectra according to Mayagaya et
al., randomized, and divided processed spectra into two groups. The first group
contained 70% of the total spectra and was used for training models. The second
group had 30% of the total spectra and was used for out-of-sample testing.

We trained a PLS ten-component model using ten-fold cross validation [132].
Even though a range of six to ten PLS components were used in previous
studies [35, 75, 119, 120, 121], we used ten PLS components after plotting the percentage of variance explained in the dependent variable against the number of PLS components (Figure 2.1). For both species, there is not much change in the percentage variance explained in the dependent variables beyond ten components.

Figure 2.1: The percentage of variance explained in the dependent variable against the number of PLS components: A) *An. gambiae* B) *An. arabiensis*
For the ANN model, we trained a feed-forward ANN with one hidden layer, ten neurons, and a linear transfer function (purelin) using Levenberg-Marquardt (damped least-squares) optimization [8]. We used actual mosquito ages as labels during training of both PLS and ANN regression models. We determined whether the trained models are over-fit by applying trained models (PLS and ANN) to estimate the ages of mosquitoes on both training (in sample) and test (out-of-sample) data sets. Normally, if the model is not over-fit, the accuracy of the model is consistent between training and test sets [19].

The accuracies of the models were determined by computing their root mean squared error (RMSE) [20, 53, 145]. We evaluated the influence of the model architecture on the model accuracy by comparing their accuracies.

When interpreting the regression models as binary classifiers, mosquitoes with an estimated age less than seven days were considered as less than seven days old, and those greater or equal to seven days were considered older than or equal to seven days old. Using Equations 2.1, 2.2, and 2.3, we computed and compared sensitivity, specificity, and accuracy between the PLS and ANN regression models interpreted as binary classifiers. Sensitivity of the model is the ability to classify mosquitoes correctly, which are older than or equal to seven days old (assumed to be positively related to malaria transmission), and specificity is the ability of the
model to classify mosquitoes correctly which are less than seven days old (assumed to be negatively related to malaria transmission) [3, 63, 130].

Let

- True Positive (TP) = Number of mosquitoes correctly predicted \( \geq 7 \) days old,
- True Negative (TN) = Number of mosquitoes correctly predicted < 7 days old,
- Positive (P) = Total number of mosquitoes greater \( \geq 7 \) days old, and
- Negative (N) = Total number of mosquitoes less than 7 days old.

Then

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{P}}, \quad (2.1)
\]

\[
\text{Specificity} = \frac{\text{TN}}{\text{N}}, \quad \text{and} \quad (2.2)
\]

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}. \quad (2.3)
\]

Training a regression model and interpreting it as a binary classifier can compromise the accuracy of the model as a classifier. This is because, while training a regression model forces the model to learn differences between actual ages of mosquitoes, direct training of a binary classifier forces the model to learn similarities between mosquitoes of the same class and only differences between two classes. Therefore, we directly trained binary classification models using ANN and PLS,
architectures and compare the accuracies with the ANN and PLS regression models interpreted as binary classifiers. In both species, we divided processed spectra (786 spectra for *An. gambiae* and 870 spectra for *An. arabiensis*) into two groups; less than seven days old and greater than or equal to seven days old. The spectra in a group with mosquitoes less than seven days old were labeled 0, and 1 for those in a group with mosquitoes greater than or equal to seven days old, and the two groups were merged. The spectra were randomized and divided into training (N = 508 for both species) and test (N = 278 for *An. gambiae* and N = 362 for *An. arabiensis*) sets. We trained a PLS ten-component model using ten-fold cross-validation [132] and a one hidden layer, ten neuron feed-forward ANN using logistic regression as a transfer function and Levenberg-Marquardt (damped least-squares) optimization for training [8, 90]. During interpretation of these models, mosquitoes scored < 0.5 were considered as less than 7 days old and those scored ≥ 0.5 as seven days old or more. Using Equations 2.1, 2.2, and 2.3, for each species, we computed specificity, sensitivity, and accuracy of the trained PLS and ANN binary classifiers and compared to the PLS and ANN regressors interpreted as the binary classifiers. We repeated the process of random splitting the dataset into training and test sets; training, testing and scoring the accuracies of trained models ten times and compare the average results, a process known as Monte Carlo cross-validation [37, 147, 148].

To test the reproducibility of our results, we applied a similar analysis to
different data sets of mosquitoes already used in other publications and freely available for re-use [62, 95, 121, 126, 127] (Figure 2.2). Tables 1.1 and 1.2 (in Chapter 1), respectively, summarize key information and number of mosquitoes per age group in each data set. Details on these data sets can be found in their respective publications.

Figure 2.2: Reproducing our analysis on different datasets (black DS = An. gambiae; green DS = An. arabiensis; red DS = Ae. aegypti and Ae. albopictus)
Despite differences in characteristics (i.e., different killing methods, different scanning instruments, and different sources of mosquitoes) of mosquitoes in our datasets (IFA-ARA and IFA-GA) and datasets 1-8 (Table 1.1), we use datasets 7 - 8 and datasets 1- 4 as independent test sets to test models trained on IFA-ARA and IFA-GA, respectively, (Figure 2.3). Normally a fair testing of a trained model should be on datasets whose samples have the same characteristics as the samples used to train the model. Here, we wanted to compare how ANN and PLS models extrapolate to datasets whose samples may have different characteristics than the samples used to train them.

Figure 2.3: ANN and PLS models trained on IFA-ARA and IFA-GA datasets were tested on independent datasets.
2.3 Results

Both PLS and ANN regression models consistently estimated the age of *An. gambiae* and *An. arabiensis* in the training and test data sets, showing that the models likely were not over-fit on these datasets during training (Figures 2.4 and 2.5). Figures 2.6 - 2.7 and Tables 2.1 - 2.3 present the performances of PLS and ANN regression models when estimating actual age of *An. gambiae* and *An. arabiensis* in the test data set and when their outputs are interpreted into two age classes, showing significant differences in accuracies of the two models (PLS vs. ANN models). ANN regression model scores significantly higher accuracy than the PLS regression model. Tables 2.4 and 2.5 represent results when the same analysis was extended to different datasets of *An. arabiensis*, *An. gambiae*, *Aedes aegypti* (infected and non-infected with *Wolbachia*) and *Aedes albopictus* already used in other publications, showing reproducibility of the results presented in Table 2.1 (ANN performing better than PLS models).

Figure 2.8 represents consistency in accuracy of PLS (A and C) and ANN (B and D) directly trained binary classifiers on estimating both training and test data sets, showing that the models likely were not over-fitted during training.

Figures 2.9 - 2.10 and Table 2.6 present the results when directly trained PLS (A and C) and ANN (B and D) binary classifiers were applied to classify ages of *An.*
Figure 2.4: PLS (A and C) and ANN (B and D) regression models, estimating actual age of training and testing samples of *An. gambiae* (A and B) and *An. arabiensis* (C and D), respectively. The results further show that in both species, irrespective of the architecture used to train the model, direct training of the binary classifier scores significantly higher accuracy, specificity, and sensitivity than the regression
Figure 2.5: Regression coefficients weights against wavelengths: A) *An. gambiae* B) *An. arabiensis*

model translated as a binary classifier (Table 2.7). This observation was not only true to our dataset but also observed when the same analysis was applied to different datasets of mosquitoes already used in other publications [62, 95, 121, 127, 128] (Tables 2.8 and 2.9).
Table 2.10 presents results when our models trained on IFA-ARA and IFA-GA were tested on an independent dataset, showing that the ANN model generally performing better than the PLS model.
2.4 Discussion

This study aimed at improving the current state of the art accuracies of the models trained using near infrared spectra to estimate the age of *An. gambiae* and *An. arabiensis*. Previous studies [35, 75, 76, 120, 121, 124] trained a regression model...
Table 2.1: **Performance analysis of PLS and ANN regression models on estimating age of An. gambiae and An. arabiensis.** Results from ten-fold Monte Carlo cross-validation.

<table>
<thead>
<tr>
<th>Species</th>
<th>Model estimation</th>
<th>Metric</th>
<th>Model architecture</th>
<th>P-value (two tail)</th>
<th>P-value (one tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>An. gambiae</td>
<td>Actual age</td>
<td>RMSE</td>
<td>3.7 ± 0.2</td>
<td>1.6 ± 0.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>83.9 ± 2.3</td>
<td>93.7 ± 1.0</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>89.0 ± 2.1</td>
<td>92.5 ± 1.6</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>75.8 ± 5.2</td>
<td>95.6 ± 1.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>An. arabiensis</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.5 ± 0.1</td>
<td>2.8 ± 0.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>80.3 ± 2.1</td>
<td>90.2 ± 1.7</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>90.5 ± 1.9</td>
<td>91.7 ± 3.3</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>60.3 ± 4.2</td>
<td>88.4 ± 3.9</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 2.2: **Mean actual age estimation of mosquitoes in out-of-sample test sets by ANN and PLS regression models.** Column ‘N’ represents the number of mosquitoes in each age group.

<table>
<thead>
<tr>
<th>Actual age</th>
<th>An. arabiensis</th>
<th>An. gambiae s.s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLS</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1.9 ± 3.2</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>5.8 ± 3.9</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>9.3 ± 3.3</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>8.7 ± 2.9</td>
<td>47</td>
</tr>
<tr>
<td>9</td>
<td>9.9 ± 3.7</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>12.2 ± 3.4</td>
<td>45</td>
</tr>
<tr>
<td>15</td>
<td>13.6 ± 4.3</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>17.3 ± 3.4</td>
<td>38</td>
</tr>
<tr>
<td>25</td>
<td>19.9 ± 6.7</td>
<td>38</td>
</tr>
</tbody>
</table>

Knowing that the selection of a model architecture often influences the
Table 2.3: Percentage of mosquitoes in each age group correctly classified when ANN and PLS regression models were interpreted as binary classifiers to classify mosquitoes into either < 7 days old or ≥ 7 days old

<table>
<thead>
<tr>
<th>Species</th>
<th>Model type</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>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>An. gambiae</td>
<td>ANN</td>
<td>100</td>
<td>100</td>
<td>94.3</td>
<td>48.8</td>
<td>97.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(29/36)</td>
<td>(25/36)</td>
<td>(33/36)</td>
<td>(26/36)</td>
<td>(34/36)</td>
<td>(29/36)</td>
<td>(36/36)</td>
<td>(29/36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>100</td>
<td>100</td>
<td>61</td>
<td>72.3</td>
<td>97.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(41/36)</td>
<td>(36/36)</td>
<td>(34/36)</td>
<td>(34/36)</td>
<td>(23/36)</td>
<td>(13/36)</td>
<td>(27/36)</td>
<td>(27/36)</td>
<td></td>
</tr>
<tr>
<td>An. arabiensis</td>
<td>ANN</td>
<td>100</td>
<td>90</td>
<td>61</td>
<td>72.3</td>
<td>97.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(41/36)</td>
<td>(36/36)</td>
<td>(34/36)</td>
<td>(34/36)</td>
<td>(23/36)</td>
<td>(13/36)</td>
<td>(27/36)</td>
<td>(27/36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>100</td>
<td>57.5</td>
<td>61</td>
<td>72.3</td>
<td>97.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(41/36)</td>
<td>(36/36)</td>
<td>(10/36)</td>
<td>(23/36)</td>
<td>(13/36)</td>
<td>(27/36)</td>
<td>(27/36)</td>
<td>(27/36)</td>
<td></td>
</tr>
</tbody>
</table>

model accuracy [91], we trained age regression models using an artificial neural network [12, 50, 58, 78, 113] and partial least squares as model architectures and compared the accuracies. ANN models achieved significantly higher accuracies than corresponding PLS regression models. As summarized in Table 2.1, ANN regression models scored an average RMSE of 1.6 ± 0.2 for An. gambiae and 2.8 ± 0.2 for An. arabiensis. The PLS regression models scored RMSE of 3.7 ± 0.2 for An. gambiae and 4.5 ± 0.1 for An. arabiensis. When both ANN and PLS regression models were interpreted as binary classifiers, ANN regression model scored accuracy, sensitivity, and specificity of 93.7 ± 1.0%, 92.5 ± 1.6%, and 95.6 ± 1.8%, respectively, for An. gambiae; 90.2 ± 1.7%, 91.7 ± 3.3% and 88.4 ± 3.9%, respectively, for An. arabiensis.
Table 2.4: Reproducibility analysis of PLS and ANN regression models for estimating the age of *An. gambiae* and *An. arabiensis* in different datasets already used in other publications. DS1 - DS8 are mosquito datasets described in Tables 1.1 and 1.2. Results are from ten-fold Monte Carlo cross-validation.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Model estimation</th>
<th>Metric</th>
<th>Model architecture</th>
<th>P-value (two tail)</th>
<th>P-value (one tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>3.07 ± 0.3</td>
<td>2.6 ± 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>73.1 ± 6.3</td>
<td>85.1 ± 1.7</td>
</tr>
<tr>
<td>(N = 223)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>80.6 ± 4.2</td>
<td>84.2 ± 2.8</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>61.3 ± 9.4</td>
<td>87.5 ± 2.2</td>
</tr>
<tr>
<td>DS2</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>2.3 ± 0.1</td>
<td>1.6 ± 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>84.7 ± 3.6</td>
<td>90.2 ± 1.9</td>
</tr>
<tr>
<td>(N = 194)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>86.7 ± 5.2</td>
<td>93.3 ± 1.7</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>81.7 ± 8.1</td>
<td>90.2 ± 2.5</td>
</tr>
<tr>
<td>DS3</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>2.4 ± 0.1</td>
<td>2.3 ± 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>91.5 ± 4.5</td>
<td>93.4 ± 2.4</td>
</tr>
<tr>
<td>(N = 201)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>94.9 ± 3.4</td>
<td>93.9 ± 1.2</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>88.4 ± 5.3</td>
<td>90.2 ± 2.2</td>
</tr>
<tr>
<td>DS4</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>3.2 ± 0.1</td>
<td>2.5 ± 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>68.4 ± 2.8</td>
<td>81.7 ± 2.3</td>
</tr>
<tr>
<td>(N = 417)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>80.7 ± 2.6</td>
<td>85.6 ± 2.4</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>51.1 ± 6.8</td>
<td>75.5 ± 2.6</td>
</tr>
<tr>
<td>DS5</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>3.4 ± 0.2</td>
<td>2.7 ± 0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>69.8 ± 1.2</td>
<td>80.1 ± 2.1</td>
</tr>
<tr>
<td>(N = 618)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>81.3 ± 2.7</td>
<td>87.9 ± 3.0</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>56.7 ± 2.6</td>
<td>75.8 ± 3.4</td>
</tr>
<tr>
<td>DS6</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>2.3 ± 0.3</td>
<td>1.8 ± 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>83.1 ± 4.1</td>
<td>90.2 ± 3.4</td>
</tr>
<tr>
<td>(N = 527)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>80.3 ± 2.5</td>
<td>92.3 ± 1.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>86.9 ± 3.8</td>
<td>87.9 ± 3.7</td>
</tr>
<tr>
<td>DS7</td>
<td>Actual age</td>
<td>RMSE</td>
<td>PLS</td>
<td>2.5 ± 0.2</td>
<td>1.8 ± 0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>76.7 ± 6.2</td>
<td>80.6 ± 2.5</td>
</tr>
<tr>
<td>(N = 279)</td>
<td>Age class</td>
<td>Sensitivity (%)</td>
<td>60.8 ± 8.3</td>
<td>71.4 ± 2.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>82.3 ± 5.8</td>
<td>88.2 ± 3.2</td>
</tr>
</tbody>
</table>
Table 2.5: Performance analysis of PLS and ANN regressers for estimating the age of *Aedes albopictus*, *Wolbachia*-free, and *Wolbachia*-infected male and female *Aedes aegypti*. DS9 - DS17 are mosquito datasets described in Tables 1.1 and 1.2. Results are from ten-fold Monte Carlo cross-validation.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Model estimation</th>
<th>Metric</th>
<th>Model architecture</th>
<th>P-value (two tail)</th>
<th>P-value (one tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS9</td>
<td>Actual age</td>
<td>RMSE</td>
<td>3.8 ± 0.2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>79.4 ± 4.6</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(N = 395)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>81.7 ± 3.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>75.9 ± 6.2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS10</td>
<td>Actual age</td>
<td>RMSE</td>
<td>5.7 ± 0.1</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>77.2 ± 2.0</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(N = 600)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>78.7 ± 2.2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>72.5 ± 7.4</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS11</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.7 ± 0.2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>80.3 ± 3.1</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(N = 233)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>87.8 ± 1.2</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>78.8 ± 6.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS12</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.0 ± 0.2</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>78.3 ± 4.1</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>(N = 229)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>82.6 ± 3.6</td>
<td>0.012</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>69.1 ± 8.6</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS13</td>
<td>Actual age</td>
<td>RMSE</td>
<td>3.7 ± 0.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>87.9 ± 4.3</td>
<td>0.023</td>
<td>0.042</td>
</tr>
<tr>
<td>(N = 277)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>89.0 ± 7.2</td>
<td>0.04</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>85.2 ± 6.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS14</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.8 ± 0.1</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>83.7 ± 3.4</td>
<td>0.02</td>
<td>0.008</td>
</tr>
<tr>
<td>(N = 284)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>96.1 ± 2.6</td>
<td>0.038</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>62.6 ± 9.7</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS15</td>
<td>Actual age</td>
<td>RMSE</td>
<td>5.0 ± 0.1</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>72.9 ± 1.5</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(N = 905)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>73.2 ± 2.3</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>68.9 ± 2.4</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS16</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.2 ± 0.2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>76.6 ± 2.8</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(N = 1113)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>78.8 ± 4.3</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>73.3 ± 1.9</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DS17</td>
<td>Actual age</td>
<td>RMSE</td>
<td>4.1 ± 0.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy (%)</td>
<td>87.6 ± 2.9</td>
<td>0.006</td>
<td>0.012</td>
</tr>
<tr>
<td>(N = 585)</td>
<td></td>
<td>Sensitivity (%)</td>
<td>89.9 ± 3.0</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>85.1 ± 4.3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Figure 2.8: The consistency in accuracies of directly trained PLS (A and C) and ANN (B and D) binary classifiers for estimating age classes of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in both training and testing sets.

The PLS regression model scored accuracy, sensitivity, and specificity of $83.9 \pm 2.3\%$, $89.0 \pm 2.1\%$, and $75.8 \pm 5.2\%$, respectively, for *An. gambiae*; $80.3 \pm 2.1\%$, $90.5 \pm 1.9\%$, and $60.3 \pm 4.2\%$, respectively, for *An. arabiensis*. 
Figure 2.9: Box plot of directly trained PLS (A and C) and ANN (B and D) binary classifiers for estimating age classes of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in out-of-sample testing sets.

The interpretation of a regression model as a binary classifier can introduce errors that compromise the accuracy of the model. We directly trained PLS and ANN binary classifiers and compared the accuracies with ANN and PLS regression models interpreted as binary classifiers. Irrespective of the model architecture, directly trained binary classifiers scored significantly higher accuracies than
Figure 2.10: The number of correct and false predictions in each estimated age class when directly trained PLS (A and C) and ANN (B and D) binary classifiers were applied to classify the ages of *An. gambiae* (A and B) and *An. arabiensis* (C and D) in testing sets. Results from ten replicates.

corresponding regression models interpreted as binary classifiers (Table 2.7). The explanation of these results could be that training a regression model and interpreting it as a binary classifier involved learning differences between multiple age groups (1, 3, 5, 7, 9, 11, 13, 15, and 20 days old for *An. gambiae* and 1, 3, 5, 7, 9, 11, 13, 15, 20, and 25 days for *An. arabiensis*) of mosquitoes, which can be
Table 2.6: Comparison of the accuracy of ANN and PLS classification models on ten replicates

<table>
<thead>
<tr>
<th>Species</th>
<th>Metric</th>
<th>Model architecture</th>
<th>P-value (two-tail)</th>
<th>P-value (one-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>PLS 93.6 ± 1.2</td>
<td>0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 99.4 ± 1.0</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>An. gambiae</td>
<td>Sensitivity (%)</td>
<td>PLS 94.4 ± 1.6</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 99.3 ± 1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity (%)</td>
<td>PLS 92.4 ± 1.9</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 99.5 ± 0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>An. arabiensis</td>
<td>Accuracy (%)</td>
<td>PLS 88.7 ± 1.1</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 99.0 ± 0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sensitivity (%)</td>
<td>PLS 95.4 ± 1.4</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 99.5 ± 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity (%)</td>
<td>PLS 75.2 ± 3.4</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN 98.3 ± 1.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

challenging for two consecutive age groups. In contrast, direct training of the binary classifier involved learning differences existing between only two age groups. During direct training of the binary classifier, the process of dividing spectra into two groups (less than seven days or seven days or more) forced a model to learn similarities instead of differences between mosquitoes of the same age class. We also observed that the directly trained ANN binary classifier scored higher accuracy than the directly trained PLS binary classifier. The ANN binary classifier scored an accuracy, sensitivity, and specificity of 99.4 ± 1.0%, 99.3 ± 1.4%, and 99.5 ± 0.7%, respectively, for An. gambiae; 99.0 ± 0.6%, 99.5 ± 0.5%, and 98.3 ± 1.3%, respectively, for An. arabiensis. The PLS binary classifier scored 93.6 ± 1.2%, 94.4 ± 1.6%, and 92.5 ± 1.9% for An. gambiae; 88.7 ± 1.1%, 95.5 ± 1.4%, and 75.2 ± 3.5% for An. arabiensis (Table 2.6).

Reproducibility of results is one of the key components when testing
Table 2.7: **Comparison of accuracies between directly trained binary classifiers and regressers interpreted as binary classifiers.** Results from ten-fold Monte Carlo cross-validation. DTBC: Directly trained binary classifier; RIBC: Regresser interpreted as binary classifier.

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>Metric</th>
<th>Model type</th>
<th>DTBC</th>
<th>RIBC</th>
<th>P-value (two tail)</th>
<th>P-value (one tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An.gambiae</em></td>
<td>PLS</td>
<td>Accuracy (%)</td>
<td></td>
<td>93.6 ± 1.2</td>
<td>83.9 ± 2.3</td>
<td>3.5 × 10⁻⁰⁵</td>
<td>1.8 × 10⁻⁰⁵</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td></td>
<td>94.4 ± 1.6</td>
<td>89.0 ± 2.1</td>
<td>1.1 × 10⁻⁰³</td>
<td>5.3 × 10⁻⁰⁴</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td></td>
<td>92.4 ± 1.9</td>
<td>75.8 ± 5.2</td>
<td>1.3 × 10⁻⁰⁴</td>
<td>6.8 × 10⁻⁰⁵</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>Accuracy (%)</td>
<td></td>
<td>99.4 ± 1.0</td>
<td>93.7 ± 1.0</td>
<td>2.3 × 10⁻¹⁹</td>
<td>1.2 × 10⁻¹⁹</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td></td>
<td>99.3 ± 1.4</td>
<td>92.5 ± 1.6</td>
<td>7.3 × 10⁻⁰⁷</td>
<td>3.7 × 10⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td></td>
<td>99.5 ± 0.7</td>
<td>95.6 ± 1.8</td>
<td>2.2 × 10⁻⁰³</td>
<td>1.1 × 10⁻⁰³</td>
</tr>
<tr>
<td><em>An.arabiensis</em></td>
<td>PLS</td>
<td>Accuracy (%)</td>
<td></td>
<td>88.7 ± 1.1</td>
<td>80.3 ± 2.1</td>
<td>6.9 × 10⁻⁰⁸</td>
<td>3.4 × 10⁻⁰⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td></td>
<td>95.4 ± 1.4</td>
<td>90.5 ± 1.9</td>
<td>2.3 × 10⁻⁰⁴</td>
<td>1.2 × 10⁻⁰⁴</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td></td>
<td>75.2 ± 3.4</td>
<td>60.3 ± 4.2</td>
<td>5.5 × 10⁻⁰⁵</td>
<td>2.8 × 10⁻⁰⁵</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>Accuracy (%)</td>
<td></td>
<td>99.0 ± 0.6</td>
<td>90.2 ± 1.7</td>
<td>1.8 × 10⁻²¹</td>
<td>8.8 × 10⁻²²</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td></td>
<td>99.5 ± 0.5</td>
<td>91.7 ± 3.3</td>
<td>3.2 × 10⁻⁰⁵</td>
<td>1.6 × 10⁻⁰⁵</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td></td>
<td>98.3 ± 1.3</td>
<td>88.4 ± 3.9</td>
<td>1.1 × 10⁻⁰⁴</td>
<td>5.3 × 10⁻⁰⁵</td>
</tr>
</tbody>
</table>

precision and accuracy of a new measurement or method [6]. We further tested the reproducibility of our analyses on different datasets of *An. gambiae, An. arabiensis, Aedes aegypti* (males and females infected and not infected with *Wolbachia*), and *Aedes albopictus*, which are already published and freely available for re-use in other studies [62, 95, 121, 126, 127].
Table 2.8: Comparison of the accuracy of ANN and PLS classification models on *An. gambiae* and *An. arabiensis* in datasets from other published studies described in Table 1.1.

<table>
<thead>
<tr>
<th>Species</th>
<th>Metric</th>
<th>Model architecture</th>
<th>P-value (two-tail)</th>
<th>P-value (one-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>PLS</td>
<td>ANN</td>
<td></td>
</tr>
<tr>
<td>DS1</td>
<td>Accuracy (%)</td>
<td>93.3 ± 5.1</td>
<td>96.7 ± 1.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 201)</td>
<td>Sensitivity (%)</td>
<td>91.3 ± 7.6</td>
<td>93.9 ± 1.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DS2</td>
<td>Accuracy (%)</td>
<td>88.0 ± 3.1</td>
<td>94.2 ± 2.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 194)</td>
<td>Sensitivity (%)</td>
<td>86.0 ± 6.2</td>
<td>96.7 ± 1.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DS3</td>
<td>Accuracy (%)</td>
<td>95.6 ± 2.9</td>
<td>98.7 ± 2.3</td>
<td>0.042</td>
</tr>
<tr>
<td>(N = 250)</td>
<td>Specificity (%)</td>
<td>90.1 ± 7.1</td>
<td>94.7 ± 1.8</td>
<td>0.005</td>
</tr>
<tr>
<td>DS4</td>
<td>Accuracy (%)</td>
<td>73.8 ± 5.6</td>
<td>83.3 ± 3.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 417)</td>
<td>Sensitivity (%)</td>
<td>78.9 ± 4.3</td>
<td>84.5 ± 2.6</td>
<td>0.003</td>
</tr>
<tr>
<td>DS6</td>
<td>Accuracy (%)</td>
<td>70.9 ± 2.2</td>
<td>80.7 ± 2.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 417)</td>
<td>Specificity (%)</td>
<td>85.6 ± 3.4</td>
<td>90.1 ± 1.1</td>
<td>0.002</td>
</tr>
<tr>
<td>DS7</td>
<td>Accuracy (%)</td>
<td>83.3 ± 4.3</td>
<td>90.4 ± 2.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 527)</td>
<td>Sensitivity (%)</td>
<td>75.7 ± 4.8</td>
<td>87.5 ± 3.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DS8</td>
<td>Accuracy (%)</td>
<td>88.5 ± 3.4</td>
<td>94.2 ± 2.9</td>
<td>0.005</td>
</tr>
<tr>
<td>(N = 279)</td>
<td>Specificity (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We found consistency in results between our datasets and different datasets of mosquitoes already published in other studies (Tables 2.4, 2.5, 2.8, and 2.9).
This consistency strengthens the assertion that ANN models score higher accuracy than PLS models.

<table>
<thead>
<tr>
<th>Species Metric</th>
<th>Model architecture</th>
<th>P-value (two-tail)</th>
<th>P-value (one-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>86.9±2.8</td>
<td>92.4±2.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>93.3±4.3</td>
<td>94.5±2.8</td>
<td>0.22</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>82.3±2.2</td>
<td>88.9±1.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 395)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS10 Sensitivity (%)</td>
<td>93.9±1.9</td>
<td>96.5±1.0</td>
<td>0.034</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>89.2±1.7</td>
<td>94.6±1.8</td>
<td>0.005</td>
</tr>
<tr>
<td>(N = 600)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS11 Sensitivity (%)</td>
<td>91.9±1.3</td>
<td>97.4±1.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>85.7±2.1</td>
<td>94.6±1.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 233)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS12 Sensitivity (%)</td>
<td>91.3±1.1</td>
<td>98.3±1.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>89.9±6.7</td>
<td>93.6±3.1</td>
<td>0.04</td>
</tr>
<tr>
<td>(N = 229)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS13 Sensitivity (%)</td>
<td>94.6±2.2</td>
<td>94.4±2.9</td>
<td>0.72</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>86.4±6.8</td>
<td>98.5±1.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 277)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS14 Sensitivity (%)</td>
<td>89.0±3.1</td>
<td>90.5±2.6</td>
<td>0.277</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>73.6±5.7</td>
<td>89.4±1.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 905)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS15 Sensitivity (%)</td>
<td>84.2±1.6</td>
<td>90.8±1.2</td>
<td>0.008</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>88.4±1.2</td>
<td>96.7±2.1</td>
<td>0.003</td>
</tr>
<tr>
<td>(N = 1113)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS16 Sensitivity (%)</td>
<td>96.0±2.3</td>
<td>95.9±2.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>71.7±3.4</td>
<td>89.8±2.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 277)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS17 Sensitivity (%)</td>
<td>95.6±1.5</td>
<td>98.3±0.4</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>80.7±6.6</td>
<td>94.6±2.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(N = 585)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.9: Comparison of the accuracies of directly trained ANN and PLS classification models on *Aedes aegypti* and *Aedes albopictus* in datasets from other published studies described in Table 1.1.
Table 2.10: Results when both regression and directly trained binary classifiers trained on IFA-GA and IFA-ARA datasets were tested on DS1 - 4 and DS7 - 8 (described in Table 2.1), respectively, as independent test sets.

<table>
<thead>
<tr>
<th>Training set</th>
<th>Test set</th>
<th>Model type</th>
<th>Metric</th>
<th>Model architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PLS</td>
</tr>
<tr>
<td>IFA-GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS1</td>
<td>Regression</td>
<td>RMSE</td>
<td>5.8</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>47.9</td>
<td>60.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>36.8</td>
<td>65.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>62.2</td>
<td>56.5</td>
</tr>
<tr>
<td>DS2</td>
<td>Regression</td>
<td>RMSE</td>
<td>5.4</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>49.5</td>
<td>69.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>26.1</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>83.5</td>
<td>77.9</td>
</tr>
<tr>
<td>DS3</td>
<td>Regression</td>
<td>RMSE</td>
<td>5.82</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>63.7</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>66.7</td>
<td>76.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>60.6</td>
<td>79.4</td>
</tr>
<tr>
<td>DS4</td>
<td>Regression</td>
<td>RMSE</td>
<td>5.7</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>56.9</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>66.4</td>
<td>77.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>48.0</td>
<td>70.8</td>
</tr>
<tr>
<td>IFA-ARA</td>
<td>DS7</td>
<td>Regression</td>
<td>RMSE</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>71.4</td>
<td>82.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>67.8</td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>75.3</td>
<td>83.2</td>
</tr>
<tr>
<td></td>
<td>DS8</td>
<td>Regression</td>
<td>RMSE</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Accuracy (%)</td>
<td>48.7</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitivity (%)</td>
<td>32.3</td>
<td>73.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity (%)</td>
<td>70.3</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Our study is not the first to observe ANN models outperforming PLS models. Besides being reproducible in different datasets, these findings also are
supported with other previous studies [12, 50, 88, 150] comparing the accuracies of
ANN and PLS models, where they report that ANN models perform better than
PLS models. The explanation of these results could be that ANN, unlike PLS,
considers both linear and unknown non-linear relationships between dependent and
independent variables [12, 50, 58]; builds independent-dependent relationships that
interpolate well even to cases that were not exactly presented by training data; and
has a self mechanism of filtering and handling noisy data during training [78, 113].
Hence, ANN models are unbiased estimators, in contrast to PLS models (Figures
2.11 and 2.12).

We also found that an ANN model extrapolates better than a PLS model
when tested on datasets whose samples have different characteristics than the
samples used to train them (Table 2.10). These results strengthen the assertion that
ANNs can filter and handle noisy data better than PLS models. Furthermore, these
results suggest that training neural networks on samples with varying characteristics
such as different killing methods, scanning instruments, and geographical regions,
might yield a model with better performance than the one presented in Table 2.10.
The only caveat with this is a need for large dataset to train the model.
Figure 2.11: Error distribution per actual age of *An. gambiae* and *An. arabiensis* when ANN and PLS regressors applied to estimate the actual ages of mosquitoes in training and test data sets, showing a uniform distribution of errors (un-biased estimating) across actual ages of mosquitoes for the ANN regressor and an un-uniform distribution of errors (biased estimating) for the PLS regressor.

2.5 Conclusion

We conclude that training both regression and binary classification age artificial neural network models yield higher accuracies than partial least squares models. Also, training a binary classifier scores higher accuracy than training a regression
Figure 2.12: Error distribution per actual age of *An. gambiae* and *An. arabiensis* when ANN and PLS regressors applied to estimate the actual ages of mosquitoes in training and test data sets, showing a uniform distribution of errors (un-biased estimating) across actual ages of mosquitoes for the ANN regressor and an un-uniform distribution of errors (biased estimating) for the PLS regressor model and interpreting it as a binary classifier. Hence, we recommend training of ANN models over PLS models to estimate age of mosquitoes, and training of binary classifier instead of training regression model and interpreting it as binary classifier.
2.6 Transition to Chapter 3

In Chapter Two, using NIR spectra collected from laboratory and semi-field reared mosquitoes, we have demonstrated that models trained using an artificial neural network architecture perform better than similar models trained using a partial least squares architecture. In Chapter 3, we apply artificial neural networks to train models that estimate parity status (egg laying status) of wild mosquitoes.
CHAPTER 3

An Autoencoder and Artificial Neural Network-based Method to Estimate Parity Status of Wild Mosquitoes from Near-infrared Spectra

This chapter [85] is published in bioRxiv as a pre-print and is under review for publication at PLOS ONE.

Abstract

**Background:** After mating, female mosquitoes need animal blood to develop their eggs. In the process of acquiring blood, they may acquire pathogens, which may cause different diseases to humans such as malaria, zika, dengue, and chikungunya. Therefore, knowing the parity status of mosquitoes is useful in control and evaluation of infectious diseases transmitted by mosquitoes, where parous mosquitoes are assumed to be potentially infectious. Ovary dissections, which currently are used to determine the parity status of mosquitoes, are very tedious and limited to very few experts. An alternative to ovary dissections is near-infrared spectroscopy (NIRS), which can estimate the age in days and the infectious state of laboratory and semi-field reared mosquitoes with accuracies between 80 and 99%. No study has tested the accuracy of NIRS for estimating the parity status of wild mosquitoes.
Methods and results: In this study, we train artificial neural network (ANN) models on NIR spectra to estimate the parity status of wild mosquitoes. We use four different datasets: *An. arabiensis* collected from Minepa, Tanzania (Minepa-ARA); *An. gambiae* collected from Muleba, Tanzania (Muleba-GA); *An. gambiae* collected from Burkina Faso (Burkina-GA); and *An. gambiae* from Muleba and Burkina Faso combined (Muleba-Burkina-GA). We train ANN models on datasets with spectra preprocessed according to previous protocols. We then use autoencoders to reduce the spectra feature dimensions from 1851 to 10 and re-train ANN models. Before the autoencoder was applied, ANN models estimated parity status of mosquitoes in Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba-Burkina-GA with out-of-sample accuracies of $81.9 \pm 2.8\%$ (N=927), $68.7 \pm 4.8\%$ (N=140), $80.3 \pm 2.0\%$ (N=158), and $75.7 \pm 2.5\%$ (N=298), respectively. With the autoencoder, ANN models tested on out-of-sample data achieved $97.1 \pm 2.2\%$, (N=927), $89.8 \pm 1.7\%$ (N=140), $93.3 \pm 1.2\%$ (N=158), and $92.7 \pm 1.8\%$ (N=298) accuracies for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba-Burkina-GA, respectively.

Conclusion: These results show that a combination of an autoencoder and an ANN trained on NIR spectra to estimate parity status of wild mosquitoes yields models that can be used as an alternative tool to estimate parity status of wild mosquitoes, especially since NIRS is a high-throughput, reagent-free, and simple-to-use technique compared to ovary dissections.
3.1 Introduction to the problem of the study

Evaluation of existing malaria control interventions such as insecticide-treated nets (ITNs) and indoor residual spraying (IRS) relies upon, among other factors, the assessment of the changes occurring in the mosquito parity structure prior to and after implementation of an intervention [25, 72, 111]. The parity status of mosquitoes corresponds with their capability to transmit *Plasmodium* parasites with an assumption that parous mosquitoes are more highly capable than nulliparous mosquitoes, as they may have accessed parasite-infected blood. A shift in the parity structure towards a population with more nulliparous mosquitoes signifies a reduction in the risk of disease transmission [25, 38, 43], as the chances that mosquitoes carry malaria parasite declines [10].

The current standard technique for estimating the parity status of female mosquitoes involves dissection of their ovaries to separate mosquitoes into those that have previously laid eggs, known as the parous group (assumed to be old and potentially infectious), and those that do not have a gonotrophic history, known as the nulliparous group (assumed to be young and non-infectious) [29]. Another standard technique also based on the dissection of ovaries determines the number of times a female mosquito has laid eggs [106]. However, both techniques are laborious, time consuming, and require skilled technicians. These technical difficulties lead to
analysis of small sample sizes that often fail to capture the heterogeneity of a mosquito population.

Near infrared spectroscopy (NIRS) technology, complimented by techniques from machine learning, has been demonstrated to be an alternative tool for predicting age, species, and infectious status of laboratory and semi-field raised mosquitoes [35, 62, 73, 75, 76, 83, 95, 120, 121, 127, 128]. NIRS is a rapid, non-invasive, reagent-free technique that requires minimal skills to operate, allowing hundreds of samples to be analyzed in a day. However, the accuracy of NIRS techniques for predicting the parity status of wild mosquitoes has not been tested. Moreover, recently, it has been reported that models trained on NIR spectra using an artificial neural network (ANN) estimate the age of laboratory-reared *An. arabiensis*, *An. gambiae*, *Aedes aegypti*, and *Aedes albopictus* with accuracies higher than models trained on NIR spectra using partial least squares (PLS) [83].

In this study, we train ANN models on NIR spectra preprocessed according to an existing protocol [75] to estimate the parity status of wild *An. gambiae* s.s and *An. arabiensis*. We then apply autoencoders to reduce the spectra feature space from 1851 to 10 and re-train ANN models. The ANN model achieved an average accuracy of 72% and 93% before and after applying the autoencoder, respectively. These results strongly suggest ANN models trained on autoencoded NIR spectra as an alternative tool to estimate the parity status of wild *An. gambiae*
and \textit{An. arabiensis}. High-throughput, non-invasive, reagent free, and simple to use NIRS analysis complements the limitations of ovary dissections.

### 3.2 Ethics approvals

Ethics approvals for collecting mosquitoes in Minepa-ARA, Burkina-GA, and Muleba-GA datasets from residents’ homes were obtained from Ethics Review Boards of the Ifakara Health Institute (IHI-IRB/No. 17-2015), the Colorado State University (approval No. 09-1148H), and the Kilimanjaro Christian Medical College (Certificate No. 781), respectively.

### 3.3 Data

We use data from wild \textit{An. arabiensis} (Minepa-ARA) collected from Minepa, a village in southeastern Tanzania (already published in [84] and available for reuse), from wild \textit{An. gambiae s.s} (Muleba-GA) collected from Muleba, northwestern Tanzania, and from wild \textit{An. gambiae s.s} collected from Bougouriba and Diarkadou-gou villages in Burkina Faso (Burkina-GA).

Mosquitoes in the Minepa-ARA and Muleba-GA datasets were captured using CDC light traps placed inside residential homes. Mosquitoes that were morphologically identified as members of the \textit{Anopheles gambiae} complex were
further processed. Prior to scanning, wild mosquitoes collected in Minepa were killed by freezing at -20°C for 20 minutes and left to re-equilibrate to room temperature for 30 minutes. Wild mosquitoes collected in Muleba were killed using 75% ethanol, dissected according to the technique described by Detinova [32] to determine their parity status, and preserved in silica gel. Mosquitoes in Minepa-ARA were dissected after scanning. Following a previous published protocol to collect spectra [75], mosquitoes in both Minepa-ARA and Muleba-GA were scanned using a LabSpec 5000 near-infrared spectrometer with an integrated light source (ASD Inc., Longmont, CO). After spectra collection, mosquitoes in Minepa-ARA were dissected to score their parity status. Then polymerase chain reaction (PCR) was conducted on DNA extracted from mosquito legs (in both Minepa-ARA and Muleba-GA) to identify species type as previously described [99]. Each mosquito was labeled with a unique identifier code linking each NIR spectrum to parity dissection and PCR information.

Data from wild *An. gambiae* s.s from Burkina Faso were published in [62] and publicly available for reuse. These mosquitoes are referred to as independent test sets 2 and 3 (ITS 2 and ITS 3) in [62]. ITS 2 has 40 nulliparous and 40 parous mosquitoes, and ITS 3 has 40 nulliparous and 38 parous mosquitoes. In this study, we combine these two datasets into one dataset and refer it as Burkina-GA. Mosquitoes in Burkina-GA (*N* = 158) were collected in 2013 in Burkina Faso from
Bougouriba and Diarkadou-gou villages using either indoor aspiration or a human-baited tent trap, and their ovaries were dissected according to the Detinova method [32]. Mosquitoes were preserved in silica gel before their spectra were collected using a LabSpec4i spectrometer (ASD Inc., Boulder, CO, USA).

3.4 Model training and testing

We trained models on four datasets, namely Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba-Burkina-GA (Muleba-GA and Burkina-GA combined). Before training models, spectra in all four datasets were pre-processed according to the previously published protocol [75] and divided into two groups (nulliparous and parous). Spectra in the nulliparous and parous groups were labeled zero and one, respectively. The two groups were then merged, randomized, and divided into a training set (75%; \(N = 927\) for Minepa-ARA, \(N = 140\) for the Muleba-GA, \(N = 158\) for Burkina-GA and \(N = 298\) for Muleba-Burkina-GA) and a test set (the remaining 25% in each dataset). On each dataset, using ten Monte-Carlo cross validations [83, 148] and Levenberg-Marquardt optimization, a one hidden layer, ten-neuron feed-forward ANN model with logistic regression as a transfer function was trained and tested in MATLAB (Figure 3.1).

Based on the accuracy of the model presented in Table 3.1 in the Results and Discussion section of this Chapter, we explored how to improve the model accuracy.
Figure 3.1: Training and testing ANN model on spectra preprocessed according to Mayagaya et al. [75]. “M” is either Minepa-ARA, Muleba-GA, Burkina-GA, or Muleba-Burkina-GA

Normally a parous class, unlike a nulliparous class, often is represented by a limited number of samples, posing a problem of data imbalance during model training. In this case, a large amount of data is required to obtain enough samples in a parous class for a model to learn and characterize it accurately. Obtaining enough data for model training is always challenging. The most common ways of dealing with the data imbalance are either to discard samples from a nulliparous class to equal the number of samples in a parous class or to bootstrap samples in a parous class [137]. However, discarding data to equalize the data distribution in two classes in the training set leaves an imbalanced test set. Also, it is this imbalanced scenario to which the model will be applied in real cases. In addition, throwing away samples,
especially from data sets with a high dimension feature space, can lead to over-fitting the model. Alternatively, for datasets with a high dimension feature space, instead of discarding data from a class with a large number of samples, feature reduction techniques are employed [137]. Feature reduction reduces the size of the hypothesis space initially presented in the original data, thereby reducing the size of data required to adequately train the model. Principal component analysis (PCA) and partial least squares (PLS) are the commonly used unsupervised and supervised feature reduction methods, respectively, especially for cases whose features are linearly related [91, 117]. Autoencoders recently are used as an alternative to PCA in cases involving both linear and non-linear relationships [24, 51, 60, 68].

An autoencoder is an unsupervised ANN that learns both linear and non-linear relationships present in data and represents them in a new reduced dimension data space (which also can be used to regenerate the original data space) without losing important information [66, 67, 103]. The autoencoder has two parts, the encoder part where an original dataset is encoded to a desired reduced feature space (encoded dataset) and the decoder part where the encoded dataset is decoded to an original dataset to determine how accurately the encoded dataset represents the original dataset. Figure 3.2 illustrates an example of an autoencoder in which an 1850-feature dataset is stepwise encoded to a 10-feature dataset. There is no
formula for the number and size of steps to take to get to a desired feature size. However, taking several steps results in losing very little information, compared with taking a single step.

Figure 3.2: **Autoencoder reducing feature space dimension**

Once an encoded feature space can reconstruct the original feature space with an acceptable accuracy, the decoder is detached, and a desired model (in our case an ANN binary classifier) is trained on the encoded feature space as shown in Fig 3.3.

Egg laying appears to be affected by both linear and non-linear relationships.

Hence, we separately train autoencoders on the Minepa-ARA, Muleba-GA,
Figure 3.3: **ANN model trained on a dataset with an encoded feature space**

Burkina-GA, and Muleba-Burkina-GA datasets to reduce spectra feature dimensions from 1851 to 10 (Figure 3.4). Table 3.1 presents accuracies of reconstructing original feature spaces from their respective encoded feature spaces.
We refer to the autoencoded Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba-Burkina-GA datasets as Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA, and Encoded-Muleba-Burkina-GA, respectively. We then train ANN models on Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA, and Encoded-Muleba-Burkina-GA (Figure 3.5).
Table 3.1: Accuracies of reconstructing original feature spaces from encoded feature spaces. MSE = mean square error.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Steps</th>
<th>Encoded-Minepa-ARA</th>
<th>Encoded-Muleba-GA</th>
<th>Encoded-Burkina-GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>Step 1</td>
<td>0.0046</td>
<td>0.0029</td>
<td>0.0031</td>
</tr>
<tr>
<td></td>
<td>Step 2</td>
<td>0.00005</td>
<td>0.0027</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td>Step 3</td>
<td>0.00008</td>
<td>0.0029</td>
<td>0.0011</td>
</tr>
</tbody>
</table>

Figure 3.5: Training and testing of ANN model on autoencoded spectra. M is either Minepa-ARA, Muleba-GA, Burkina-GA, or Muleba-Burkina-GA dataset.

Finally, we used the Encoded-Burkina-GA and the Encoded-Muleba-GA datasets as independent test sets to test accuracies of ANN models trained on the Encoded-Muleba-GA dataset and on the Encoded-Burkina-GA dataset, respectively (Figure 3.6A and B).
In the next section, we present results and discuss them.

### 3.5 Results and discussion

In this Chapter, we demonstrated that near-infrared spectroscopy (NIRS) can estimate accurately the parity status of wild collected *An. arabiensis* and *An. gambiae* s.s. Referring to the published results in [83] (ANN models achieve higher accuracies than PLS models), we trained and tested an ANN model on NIRS spectra in four different datasets pre-processed according to a previous published protocol [75]. The model achieved accuracies between 55.9 and 81.9% (Table 3.2, Figures 3.7 and 3.8).
Table 3.2 further presents various metrics to score performance of our classifiers, namely sensitivity, specificity, precision, and area under the receiver operating characteristic (ROC) curve (AUC). We calculated sensitivity, specificity, precision and accuracy of the model using Equations 3.1, 3.2, 3.3, and 3.4, respectively [3, 63, 115, 130].

Let

- True Positive (TP) = Number of mosquitoes correctly classified as parous,
- False Positive (FP) = Number of mosquitoes wrongly classified as parous,
- True Negative (TN) = Number of mosquitoes correctly classified as nulliparous
- Positive (P) = Total number of mosquitoes in test set that are parous, and
- Negative (N) = Total number of mosquitoes in test set that are nulliparous.

Then

\[
\text{Sensitivity} = \frac{TP}{P}, \quad (3.1)
\]

\[
\text{Specificity} = \frac{TN}{N}, \quad (3.2)
\]

\[
\text{Accuracy} = \frac{TP + TN}{P + N}, \quad (3.3)
\]

\[
\text{Precision} = \frac{TP}{TP + FP}. \quad (3.4)
\]
Figure 3.7: Box plots showing results when ANN models trained on 75% of spectra before the autoencoder was applied and tested on the remaining spectra (25%) (out-of-sample testing). A, B, C, and D represent results for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba & Burkina-GA (mosquitoes in the Muleba-GA and the Burkina-GA datasets combined) datasets, respectively.

Sensitivity (also known as recall) is the percentage of correctly predicted parous mosquitoes, specificity is the percentage of correctly predicted nulliparous mosquitoes [83], and precision is the proportion of true parous mosquitoes out of all mosquitoes estimated by the model as parous [115].
Figure 3.8: ROC curves (AUCs presented in the last row of Table 3.2) showing results when ANN models trained on 75% of spectra before the autoencoder was applied and tested on the remaining spectra (25%) (out-of-sample testing). A, B, C, and D represent results for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba & Burkina-GA (mosquitoes in Muleba-GA and Burkina-GA datasets combined) datasets, respectively.

We presented both sensitivity and precision because different scholars prefer one metric to another, especially for cases with imbalanced data [115]. AUC was computed from the receiver operating characteristic (ROC) curve shown in Figure 3.8 generated by plotting the true parous rate against the false parous rate at different threshold settings.
Table 3.2: Performance of an ANN model trained on 75% of mosquito spectra with 1851 features (before autoencoder) and tested on the remaining 25% spectra (out of the sample testing). Minepa-ARA (Nulliparous = 656, Parous = 271), Muleba-GA (Nulliparous = 119, Parous = 21) Burkina-GA (Nulliparous = 80, Parous = 78), Mu-Bu-GA = Muleba-Burkina-GA.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Precision (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minepa-ARA</td>
<td>81.9 ± 2.8</td>
<td>79.7 ± 3.2</td>
<td>86.0 ± 1.6</td>
<td>74.3 ± 3.4</td>
<td>77.2</td>
</tr>
<tr>
<td>(N=927)</td>
<td>68.7 ± 4.8</td>
<td>37.8 ± 6.6</td>
<td>80.1 ± 2.7</td>
<td>31.3 ± 5.2</td>
<td>55.9</td>
</tr>
<tr>
<td>Muleba-GA</td>
<td>80.3 ± 2.0</td>
<td>76.5 ± 2.1</td>
<td>88.3 ± 2.3</td>
<td>77.8 ± 1.8</td>
<td>83.6</td>
</tr>
<tr>
<td>(N=140)</td>
<td>65.0 ± 4.8</td>
<td>37.0 ± 6.6</td>
<td>81.0 ± 2.7</td>
<td>30.0 ± 5.2</td>
<td>52.0</td>
</tr>
<tr>
<td>Burkina-GA</td>
<td>75.7 ± 2.5</td>
<td>70.2 ± 3.1</td>
<td>77.6 ± 2.9</td>
<td>68.8 ± 3.2</td>
<td>76.4</td>
</tr>
<tr>
<td>(N=158)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mu-Bu-GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=298)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A higher AUC is interpreted as higher predictivity performance of the model [40, 108]. The ROC curve normally presents the performance of the model at different thresholds (cut-off points), providing more information on the accuracy of the classifier [40, 108]. Table 3.3 provides confusion matrices from the last (tenth) Monte-Carlo cross validation showing model accuracy in absolute values.

We hypothesized that results presented in Tables 3.2 and 3.3, and in Figures 3.7 and 3.8 were influenced by the size of a dataset used to train the model. The model that was trained on a dataset with a relatively larger number of mosquitoes, especially in the parous class, performed better than the model trained on the dataset with fewer mosquitoes.
Table 3.3: Confusion matrices showing accuracies of the models in absolute values when the models were trained on spectra before feature reduction by autoencoder. A) Minepa-ARA, B) Muleba-GA, C) Burkina-GA and D) Muleba-Burkina-GA. Results from the last Monte-Carlo cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Actual Parity</th>
<th>Estimates</th>
<th>Nulliparous</th>
<th>Parous</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Nulliparous</td>
<td>165</td>
<td>17</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>31</td>
<td>61</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>196</td>
<td>78</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Nulliparous</td>
<td>28</td>
<td>4</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>8</td>
<td>3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>36</td>
<td>7</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Nulliparous</td>
<td>20</td>
<td>6</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>4</td>
<td>18</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24</td>
<td>24</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Nulliparous</td>
<td>46</td>
<td>9</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>14</td>
<td>22</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>60</td>
<td>31</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

The current standard preprocessing technique [75] leaves a mosquito spectrum with an 1851-dimensional feature space. Mathematically, binary inputs with a 1851-dimensional feature space present $2^{(1851)}$ hypothesis space dimensions for the model to learn [136, 34, 33]. Successful learning of such hypothesis space dimensions requires many data points (mosquitoes in our case). Finding enough wild mosquitoes, especially parous mosquitoes, for a model to learn such a hypothesis space is expensive and time consuming. Feature reduction is an alternative to overcome this, as it reduces the hypothesis space dimension initially presented by the original data, hence lowering the number of data required to train the model efficiently. Techniques such as principal component analysis...
(PCA) [91, 117], partial least squares (PLS) [1, 45, 91], singular value decomposition (SVD) [45, 51, 64], and autoencoders can reduce the feature space to a size that can be learned by the available data without losing important information. PCA, PLS, and SVD are commonly used when features are linearly dependent [91, 117], otherwise, an autoencoder, which can be thought as a nonlinear version of PCA, is used [24, 51, 60, 68].

Therefore, we applied an autoencoder as illustrated in Figure 3.2 to reduce the spectra feature space from 1851 features to 10 features (Table 3.1 presents the accuracies of reconstructing original feature spaces from the encoded (reduced) feature spaces), cutting down hypothesis space dimensions from $2^{2^{\text{(1851)}}}$ to $2^{2^{\text{(10)}}}$, and re-trained ANN models (Figures 3.3 and 3.5). As presented in Tables 3.4 - 3.5 and in Figures 3.9 - 3.10, the accuracy of the model improved from an average of 72% to 93%, suggesting an ANN model trained on autoencoded NIR spectra as an appropriate tool to estimate the parity status of wild mosquitoes.

We further applied a model trained on the Muleba-GA dataset to estimate the parity status of mosquitoes in the Burkina-GA dataset and a model trained on the Burkina-GA dataset to estimate the parity status of mosquitoes in the Muleba-GA dataset. Here we wanted to test how the model performs on mosquitoes from different cohorts.
Table 3.4: Performance of an ANN model trained on 75% of the encoded mosquito spectra (10 features) and tested on the remaining 25% of the encoded mosquito spectra. Minepa-ARA (Nulliparous = 656, Parous = 271), Muleba-GA (Nulliparous = 119, Parous = 21) Burkina-GA (Nulliparous = 80, Parous = 78), Mu-Bu-GA = Muleba-Burkina-GA.

<table>
<thead>
<tr>
<th></th>
<th>Minepa-ARA (N=927)</th>
<th>Muleba-GA (N=140)</th>
<th>Burkina-GA (N=158)</th>
<th>Mu-Bu-GA (N=298)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>97.1 ± 2.2</td>
<td>89.8 ± 1.7</td>
<td>93.3 ± 1.2</td>
<td>92.7 ± 1.8</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>94.9 ± 1.6</td>
<td>70.1 ± 2.3</td>
<td>91.7 ± 1.9</td>
<td>88.2 ± 2.9</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>98.6 ± 1.3</td>
<td>96.9 ± 1.2</td>
<td>96.4 ± 1.6</td>
<td>94.7 ± 2.1</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>93.7 ± 2.4</td>
<td>62.5 ± 3.2</td>
<td>91.3 ± 1.4</td>
<td>93.1 ± 2.5</td>
</tr>
<tr>
<td>AUC (%)</td>
<td>96.7</td>
<td>91.5</td>
<td>93.1</td>
<td>94.9</td>
</tr>
</tbody>
</table>

Table 3.5: Confusion matrices showing accuracies of the models in absolute values when the models were trained on spectra after feature reduction by autoencoder. A) Minepa-ARA, B) Muleba-GA, C) Burkina-GA and D) Muleba-Burkina-GA. Results from the last Monte-Carlo cross validation.

<table>
<thead>
<tr>
<th>Actual Parity</th>
<th>Estimates</th>
<th>Nulliparous</th>
<th>Parous</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>192</td>
<td>7</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>4</td>
<td>71</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>196</td>
<td>78</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>33</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>36</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>22</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>2</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>58</td>
<td>4</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Parous</td>
<td>2</td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>60</td>
<td>31</td>
<td>91</td>
</tr>
</tbody>
</table>

As presented in Table 3.6, the model performed with accuracies of 68.6% and 88.3%, respectively, showing a model trained on the Burkina-GA dataset.
Figure 3.9: Box plots showing results when ANN models trained on 75% of encoded spectra in datasets were tested on the remaining encoded spectra (25%). A, B, C, and D represent results for Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA, and Encoded-Muleba & Burkina-GA (mosquitoes in Encoded-Muleba-GA and Encoded-Burkina-GA datasets combined) datasets, respectively.

extrapolates well to mosquitoes from a different cohort than a model trained on the Muleba-GA dataset. A possible explanation of the results shown in Table 3.6 could be that, unlike for the Burkina-GA dataset, the number of parous mosquitoes (N = 21) in the Muleba-GA dataset was not representative enough for a model to learn important characteristics that extrapolate to mosquitoes in a cohort other than the
Figure 3.10: ROC curves (AUCs presented in the last row of Table 2) showing results when ANN models trained on 75% of encoded spectra were tested on the remaining encoded spectra (25%). A, B, C, and D represent results for Encoded-Minepa-ARA, Encoded-Muleba-GA, Encoded-Burkina-GA and Encoded-Muleba & Burkina-GA (mosquitoes in Encoded-Muleba-GA and Encoded-Burkina-GA datasets combined) datasets, respectively.

one used to train the model. Although the Muleba-GA model had poor sensitivity as presented in Table 3.6, the Burkina-GA model results still suggest that the ANN model trained on acceptable number of both encoded parous and nulliparous can be applied to estimate parity status of mosquitoes from different cohorts other than the one used to train the model.
Table 3.6: Independent testing of ANN models trained on the Muleba-GA and on the Burkina-GA encoded datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>68.6</td>
<td>88.3</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>26.5</td>
<td>86.1</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>94.4</td>
<td>92.2</td>
</tr>
</tbody>
</table>

3.6 Conclusion

These results strongly suggest applying autoencoders and artificial neural networks to NIRS spectra as an appropriate complementary method to ovary dissections to estimate the parity status of wild mosquitoes. The high-throughput nature of near-infrared spectroscopy provides a statistically acceptable sample size to draw conclusions on parity status of a particular wild mosquito population. Before this method can be used as a stand-alone method to estimate parity status of wild mosquitoes, we suggest repeating of the analysis on different datasets with much larger mosquito sample sizes to test the reproducibility of the results. Hence, with the results presented in this manuscript, we recommend complementing ovary dissection with ANN models trained on NIRS spectra with their feature reduced by an autoencoder to estimate the parity status of a wild mosquito population.
3.7 Transition to Chapter 4

In addition to the importance explained in Chapter Three, parity status also is used to estimate the age of mosquitoes, where mosquitoes that have laid eggs (parous mosquitoes) are assumed to be older than mosquitoes that have not laid eggs (nulliparous mosquitoes). This interpretation is limited, as mosquitoes lay eggs after accessing blood. Hence, a mosquito can be old without laying eggs and can be young and has laid eggs. Training models that estimate the age in days of wild mosquitoes require samples of wild mosquitoes with known age in days labels. Unfortunately, it is very expensive and time consuming to gather such samples. Alternatively, models are trained on semi-field raised mosquitoes to estimate the age in days of wild mosquitoes. This practice is appropriate only if there is no significant difference between NIR spectra collected from semi-field raised and from wild mosquitoes, but no study has been done to validate this generalization. Hence, in Chapter Four, we apply clustering techniques to determine if there is any significant difference between spectra collected from semi-field and from wild mosquitoes.
CHAPTER 4

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

This chapter is adapted from [84], as published in *PLoS One*

Abstract

**Background:** Near Infrared Spectroscopy (NIRS) is a high throughput technique that measures light absorbance of biological samples and classifies the age of lab-reared mosquitoes as younger or older than seven days with an accuracy between 80 - 99%. For NIRS to estimate ages of wild mosquitoes, a sample of wild mosquitoes with known age in days is required to train and test the model. Methods such as mark-release-recapture and molecular analyses, which would provide actual ages in days of wild mosquitoes, are very difficult, tedious, time inefficient, expensive, and restricted due to their ethical implications. Alternatively, a model trained on spectra from semi-field reared mosquitoes where age in days is known can be applied to estimate the age of wild mosquitoes, but this would be appropriate only if spectra collected from semi-field reared and wild mosquitoes are similar.

**Methods and results:** We performed \( k \)-means \((k = 2)\) cluster analysis on a mixture of spectra collected from semi-field reared and wild *An. arabiensis* to
determine if there is any significant difference. While controlling the age of mosquitoes, we found two clusters with no significant difference in distribution of spectra collected from semi-field and wild mosquitoes \((p = 0.25)\). We repeated the analysis using hierarchical clustering, and similarly, no significant difference was observed \((p = 0.13)\).

**Conclusion:** We find no difference between spectra collected from semi-field reared and wild mosquitoes of the same age and species. The results strengthen and support the on-going practice of applying models trained on spectra collected from semi-field reared mosquitoes, especially first-generation semi-field reared mosquitoes.

4.1 **Introduction to mosquito age estimation**

The age of wild mosquitoes commonly is estimated by dissection of ovaries to determine their egg laying history. Mosquitoes found to have laid eggs are assumed to be older than those without an egg laying history. While generally valid, this assumption has challenges, as mosquitoes can be old without an egg laying history or young and have laid eggs. Dissection also is laborious, difficult, and limited to a few experts.

Mosquito ovary dissections [11, 30, 31, 32] may be complemented with near infrared spectroscopy (NIRS). NIRS is a high throughput technique, measuring the
energy absorbed by biological samples [13, 15, 131]. NIRS has been applied to identify species of insects infecting stored grains [36]; to differentiate between species and subspecies of termites [2]; to age-grade houseflies [101], stored grain pests [102], and biting midges [110]; to estimate the age and identify species of morphologically indistinguishable laboratory reared and semi-field raised Anopheles gambiae and Anopheles arabiensis [75, 120]; to detect and identify two strains of Wolbachia pipiens (wMelPop and wMel) in male and female laboratory-reared Aedes aegypti mosquitoes [127]; and to classify the age of male and female wild-type and Wolbachia-infected Aedes aegypti [128].

Several studies report that NIRS can classify the age of lab-reared and semi-field mosquitoes into either less than or greater than seven days old with an accuracy between 80 - 99% [75, 81, 82, 83, 95, 120]. Semi-field mosquitoes are offspring from wild caught females, raised within a large field cage (21 x 9. 1 x 7.1 m) that mimics natural mosquito habitats [93]. The ability of NIRS to estimate the age of laboratory and semi-field raised mosquitoes is a prerequisite for accurately predicting the age of wild mosquito samples.

However, it is challenging to develop a NIRS model using a sample of wild mosquitoes that can estimate the age of wild mosquitoes, as it is difficult and expensive to obtain wild mosquitoes of a known age in days with which to train and validate the model. As an alternative, models trained on spectra from semi-field
reared mosquitoes are applied to estimate the age of wild mosquitoes [62, 128], but no studies validate this generalization. Based on the NIR spectra alone, can we distinguish semi-field reared mosquitoes from wild mosquitoes? If semi-field reared and wild mosquitoes produce similar spectra, the practice of applying models trained on semi-field reared mosquitoes to estimate age of wild mosquitoes is appropriate.

In this Chapter, we first performed $k$-means cluster analysis on three different datasets generated from a spectra dataset obtained after mixing spectra at 1851 frequencies collected from: 863 semi-field reared An. Arabiensis of ages 1, 3, 5, 7, 9, 11, 15, 20, and 25 days post emergence, with at least 80 mosquitoes in each age group; and spectra collected from 927 wild An. arabiensis. We then tested the reproducibility of the results from $k$-means using hierarchical clustering. We tested the null hypothesis that there is no significant difference between the spectra collected from semi-field reared and those from wild mosquitoes when other factors are equal. We find no difference in spectra collected from semi-field reared and wild mosquitoes of the same age and species. The results strengthen the idea, which is already being practiced [62, 128], of training the model to estimate age of wild mosquitoes using spectra collected from semi-field reared mosquitoes.
4.2 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) [5, 57, 114, 138, 143]. 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 [143]. Ignoring labels from objects allows assigning of objects into groups using the objects’ features and not objects’ labels. If we translate this definition to our problem, it means that during analysis, we do not label spectra as semi-field or wild. We only provide entire spectra (absorbances) at 1851 wavelengths and partition the spectra into two groups, depending only on their absorbances and not their labels (source of a mosquito or age). If spectra collected from semi-field reared and those from wild are different, we expect them to be grouped into different clusters; otherwise they should distribute equally in the formed clusters. In this study, we applied $k$-means cluster analysis and tested the reproducibility of the results using hierarchical clustering.

$K$-means cluster analysis, also known as Lloyd’s algorithm [69], starts by arbitrarily choosing cluster centers known as centroids, depending on the number of clusters needed. In our case, we needed two clusters, as we need to determine if there is any significant difference between spectra collected from semi-field reared
and wild mosquitoes, 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 to assign each object to its closest centroid. There are different ways to compute distance, but this algorithm uses squared Euclidean distance [149]. The average distance of objects assigned to each centroid is computed. The process repeats by selecting new centroids and reassigning objects until the average distance to the centroids is minimized. More about $k$-means clustering can be found in [5, 69].

When the clusters are formed, the next step is to evaluate their quality. Evaluating the quality of the formed clusters is one of the key steps in cluster analysis and is done in many studies [42, 46, 59, 109, 116] by computing the silhouette coefficient (SC) of the cluster. SC is defined as the measure of how objects in the same cluster are similar and different from the objects in the other clusters [114, 140]. The SC of the cluster is an average of all SCs of objects in that cluster, computed using Equation 4.1.

Let

$$s(o) = \text{Silhouette coefficient of a single object 'o'},$$

$$a(o) = \text{Average distance of object 'o' to the other objects in its cluster},$$

$$b(o) = \text{Average distance of object 'o' to other objects in the nearest cluster}.$$
\[ s(o) = \frac{[b(o) - a(o)]}{\max(a(o), b(o))}. \]  

(4.1)

The lower the ‘a’ value the better, and the higher the ‘b’ value the better.

SC values ranges from -1 to +1, where +1 indicates that an object is well matched to objects in its own cluster and poorly matched to objects in neighboring clusters [114]. If most objects in the cluster have high SC, then the clustering is appropriate; otherwise, (lower SC) the clustering is inappropriate. Since SC of -1 and +1 are extreme values, the interpretation of high or low for SC values between - 1 and +1 can be subjective. The interpretation of SC due to Struyf et al. [138] summarized in Table 4.1 often is used by studies [23, 61, 98, 138, 151] involving cluster analysis.

<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>( \leq 0.25 )</td>
<td>No substantial cluster has been found</td>
</tr>
</tbody>
</table>
Hierarchical clustering groups data objects into a hierarchy or tree of clusters [55]. We built the hierarchical tree using an agglomerative method (bottom-up strategy) [55]. The agglomerative method starts by treating individual objects as clusters and then iteratively merges them into larger clusters based on their similarities [55]. Hierarchical clustering often is believed to form higher quality clusters than $k$-means, but it is limited because of its quadratic time complexity [133]. An advantage of using $k$-means is that its time complexity is linear in the number of objects, but it is thought to produce lower quality clusters [133]. Applying both $k$-means and agglomerative hierarchical approaches takes advantage of the strengths in both methods. In addition to forming quality clusters, hierarchical clustering iteratively builds different levels of clusters from clusters consisting of individual objects to one large cluster, providing a platform to analyze in detail how mosquitoes distribute in different levels of clusters in the hierarchy.

4.3 Data collection and processing

We used semi-field reared *Anopheles arabiensis* mosquitoes of ages 1, 3, 5, 7, 9, 11, 15, 20, and 25 days post emergence with at least 80 mosquitoes in each age group, from the Ifakara Health Institute semi-field system. *An. arabiensis* mosquitoes are reared in 35cm x 35cm cages in a semi-field system (SFS) [93] under ambient temperature and light-dark cycles. The humidity is artificially increased to
approximately 80% during the dry season (May - October). Adult mosquitoes are daily given a 10% glucose solution and a blood meal twice per week via human arm (Ethical clearance No. IHRDC/EC4/CL.N96/2004). The insectary keeps records of mosquitoes from egg laying to adult emergence, and the cages are labeled in such a way that mosquito ages are easily identified.

Wild *An. arabiensis* mosquitoes were collected using CDC light traps [139] in Minepa, a village located in south-eastern Tanzania. The traps were set in selected houses in the evening and collected the next morning. Live *Anopheles gambiae* complex mosquitoes were sorted from other mosquitoes from the traps and put in a small cage with cotton dipped in 10% sugar solution at the top of the cage. The sorted *Anopheles gambiae* complex mosquitoes were transported to the Ifakara Health Institute laboratory.

Before scanning, both semi-field reared and wild mosquitoes were killed by freezing for 20 minutes and left to equilibriate to room temperature for 30 minutes. Spectra were collected using a LabSpec 5000 NIR spectrometer (ASD a Panalytical company, Boulder, Colorado) as previously described [75]. After scanning, wild mosquitoes were dissected to determine their egg laying history, followed by polymerase chain reaction (PCR) to identify species type [99]. Only spectra from wild mosquitoes identified as *Anopheles arabiensis* were used for analysis. Spectra were pre-processed as previously described by Mayagaya et al. Our final dataset
contained spectra from 863 semi-field reared mosquitoes and 927 wild-caught mosquitoes at frequencies 500-2350 nm.

### 4.4 Cluster analysis and results

After spectra pre-processing, we ignored associated labels (semi-field or wild) identifying the source of mosquitoes and performed \( k \)-means cluster analysis in four different ways.

**\( K \)-means approach one:** We mixed all 863 spectra collected from semi-field reared *An. arabiensis* of ages 1, 3, 5, 7, 9, 11, 15, 20, and 25 days after emergence, with at least 80 mosquitoes in each age group, and all 927 spectra collected from wild *An. Arabiensis* and performed \( k \)-means cluster analysis on the entire data set (using 1851 absorbances at frequencies between 500 – 2350 nm) using the cluster analysis tool in MATLAB. Following the multidimensional nature of the formed clusters, it is not possible to represent the formed cluster with all absorbances in the spectra in two dimensions. Instead, for illustrative purposes, Figure 4.1 represents the formed clusters plotted using spectra according to their absorbance at two different wavelengths, 500 and 501nm (these two absorbances at 500nm and 501nm should not be confused as the only absorbances used for clustering, we used all the absorbances in the spectra during cluster analysis). Similar displays were generated using absorbances at different frequencies, and the
patterns of the displays were similar. Figure 4.1 shows that there are two clusters, despite some overlapping of spectra (objects) in both clusters.

**Figure 4.1:** Two-dimensional plot of clusters using absorbances at 500 nm and 501 nm, when the age of mosquitoes was not controlled

Figure 4.2, panel A, represents the SC of each spectrum (object) in its cluster, showing an average SC of 0.63 and 0.75 for clusters one and two, respectively. By the SC interpretation in Table 4.1, the clusters shown in Figure 4.1 are reasonable and strong, respectively.
Figure 4.2: Box plots of silhouette coefficients and bar graphs of percentage of mosquitoes, respectively, showing the quality and distribution of laboratory-reared and wild mosquitoes in clusters after k-means analysis. Panels A and B, age of mosquitoes was not controlled \((p = 0.01)\); C and D, age structure of laboratory-reared mosquitoes was controlled to match the published age structure of wild mosquitoes \((p = 0.57)\); E and F, laboratory-reared mosquitoes at 3, 5, and 25-day old were not included in the analysis \((p = 0.26)\). \(P\) stands for \(p\)-value, and \(N\) for the number of mosquitoes.

After finding the quality of the formed clusters to be reasonable and strong, respectively, a contingency table was generated, and a \(X^2\) statistical test was
performed to determine if there is a significant difference in distribution of semi-field reared and wild mosquitoes in the two clusters. That is, do the two clusters capture the sources of the mosquitoes? Figure 4.2, panel B, and Table 4.2, row $A_k$, present the results, showing a significant difference ($p = 0.01$) in the distribution of both semi-field reared and wild mosquitoes in the clusters. Cluster one has more semi-field reared mosquitoes, while cluster two has more wild mosquitoes.

The results from the $X^2$ analysis show that there is a significant difference between the spectra of wild and semi-field reared mosquitoes. However, we suspect the clustering might be age-related, and not source-related for the following reasons: several studies [75, 83, 120] show that spectra can be used to classify mosquitoes into two age classes (less than seven days against greater or equal to seven days old), implying that the age of a mosquito should not be ignored as a factor contributing the formation of two clusters; the age structure of wild mosquito populations generally follows an exponential decay curve (contain more young mosquitoes than old mosquitoes) [11, 17, 30, 142]. If our wild mosquito data have such an age distribution, and since the semi-field reared mosquitoes have a uniform age distribution by experimental design, there is high chance that clustering in the first approach was influenced by this age structure difference between the two data sets (wild and semi-field reared mosquitoes).
Table 4.2: Number and type of mosquitoes in clusters when *k*-means and hierarchical clustering were applied to spectra with: Age of mosquitoes not controlled (*A*<sub>k</sub> and *A*<sub>*h*</sub>, respectively); Age structure of semi-field reared mosquitoes controlled to match the published age structure of wild mosquitoes (*B*<sub>k</sub> and *B*<sub>*h*</sub>, respectively) and; semi-field reared mosquitoes at age 3, 5, and 25-day old not included in the analysis (*C*<sub>k</sub> and *C*<sub>*h*</sub>, respectively). \(X^2\) = computed chi-square

<table>
<thead>
<tr>
<th>Spectra data</th>
<th>Clustering technique</th>
<th>Cluster</th>
<th>Number of semi-field mosquitoes</th>
<th>Number of wild mosquitoes</th>
<th>Total</th>
<th>(X^2)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A</em>&lt;sub&gt;k&lt;/sub&gt;</td>
<td><em>K</em>-means</td>
<td>1</td>
<td>498</td>
<td>476</td>
<td>974</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>365</td>
<td>451</td>
<td>816</td>
<td>7.28</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>863</td>
<td>927</td>
<td>1790</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A</em>&lt;sub&gt;<em>h</em>&lt;/sub&gt;</td>
<td>Hierarchical</td>
<td>1</td>
<td>365</td>
<td>479</td>
<td>844</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>498</td>
<td>448</td>
<td>946</td>
<td>15.78</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>863</td>
<td>927</td>
<td>1790</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B</em>&lt;sub&gt;k&lt;/sub&gt;</td>
<td><em>K</em>-means</td>
<td>1</td>
<td>167</td>
<td>160</td>
<td>327</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>139</td>
<td>146</td>
<td>285</td>
<td>0.32</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>306</td>
<td>306</td>
<td>612</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>B</em>&lt;sub&gt;<em>h</em>&lt;/sub&gt;</td>
<td>Hierarchical</td>
<td>1</td>
<td>58</td>
<td>61</td>
<td>119</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>248</td>
<td>245</td>
<td>493</td>
<td>0.094</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>306</td>
<td>306</td>
<td>612</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C</em>&lt;sub&gt;k&lt;/sub&gt;</td>
<td><em>K</em>-means</td>
<td>1</td>
<td>337</td>
<td>495</td>
<td>832</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>261</td>
<td>432</td>
<td>693</td>
<td>1.28</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>598</td>
<td>927</td>
<td>1525</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C</em>&lt;sub&gt;<em>h</em>&lt;/sub&gt;</td>
<td>Hierarchical</td>
<td>1</td>
<td>132</td>
<td>175</td>
<td>307</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>466</td>
<td>752</td>
<td>1218</td>
<td>2.31</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>598</td>
<td>927</td>
<td>1525</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In our second *k*-means approach, we explore possible age-dependencies that may influence our clustering.

*K*-means approach two: To test whether the results in the first approach may have been influenced by age, we repeated the *k*-means analysis, this time controlling the ages of mosquitoes. Lacking age in days labels for spectra collected
from wild mosquitoes, we controlled age of mosquitoes in three different ways: First, by transforming the initial uniform age structure of semi-field reared mosquitoes to fit the published age structure (exponential decay curve) of wild mosquito populations [4, 11, 17, 70, 32, 142]. We imitate the population of semi-field reared mosquitoes with 102 one-day-old mosquitoes (based on the number of one day old in the data set) and compute the composition of other ages in the population using a published daily survival rate of 0.83 [70]. The computed number of semi-field reared mosquitoes with ages other than one day old required to form an exponential decay distribution was randomly selected from a stratified by age original semi-field reared mosquito data set. There are a number of assumptions when simulating the exponential age distribution of mosquitoes [70, 71]. The main assumptions for this simulation were: no addition of other mosquitoes into the population; the probability of a mosquito surviving one day is constant in all age classes. This process yielded a total of 306 semi-field reared mosquitoes in an imitated population. More on how to simulate the age structure of wild mosquito populations can be found at [17, 70, 71]. Figure 4.3 presents age composition in a population of semi-field reared mosquitoes selected to imitate an exponential age decay curve.
We then randomly selected 306 spectra collected from wild mosquitoes to match the number of semi-field reared mosquitoes in the selected population, mixed the two populations (selected semi-field reared to form an exponential decay distribution and randomly selected wild), and repeat $k$-means cluster analysis as in approach one. The formed clusters scored SC of 0.74 and 0.64, showing the cluster qualities to be strong and reasonable, respectively (Figure 4.2, panel C).
distribution of mosquitoes in the clusters was independent of the source of mosquitoes (Figure 4.2, panel D and Table 4.2, row $B_k$). The outcome strengthens our hypothesis that age influenced the previous clustering.

Second, we randomly selected 80 spectra collected from wild mosquitoes and maintained them for the rest of the analysis, while changing the age of the semi-field reared mosquitoes. We mixed 80 spectra of one-day-old semi-field reared An. *Arabiensis* and 80 randomly selected spectra from wild An. *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 semi-field reared mosquitoes, while keeping the spectra from wild An. *Arabiensis* unchanged (the same 80 randomly selected). Figure 4.4 illustrates the process.

The source of mosquitoes influenced the formation of clusters when clustering involved semi-field reared mosquitoes at ages 3, 5, and 25 days old (Table 4.3). For the remaining age groups, clustering was independent of the source of mosquitoes. The likely explanation for these results is that a majority of the wild mosquitoes collected could have been newly emerged but not too old.

Third, as represented in Table 4.3, only semi-field reared mosquitoes that were 3, 5, and 25 days old clustered differently from the randomly selected sample of wild mosquitoes. We hypothesized that the wild mosquitoes in the data set could have been newly emerged but not too old, causing few or none of them to be 3, 5, or
Figure 4.4: Illustration of the second method used to control number of mosquitoes per age during $k$-means approach two clustering

25 days old. Hence, creating age structure differences between semi-field reared and wild mosquito populations used in the first approach. Therefore, spectra associated with mosquitoes that are 3, 5, and 25 days old were excluded from the semi-field reared data set to determine if they influenced results in the first approach.
Table 4.3: Number of mosquitoes in clusters when 80 spectra collected from wild mosquitoes were randomly selected and maintained for the rest of the analysis, while changing the age of the semi-field reared mosquitoes. Av.SC: Average silhouette coefficient; $X^2$: Chi square

<table>
<thead>
<tr>
<th>Age</th>
<th>Cluster</th>
<th>Number of semi-field</th>
<th>Number of wild</th>
<th>Total</th>
<th>Av.SC</th>
<th>$X^2$</th>
<th>$p$-value</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>0.40</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>46</td>
<td>42</td>
<td>88</td>
<td>0.77</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>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>33</td>
<td>46</td>
<td>79</td>
<td>0.64</td>
<td>4.23</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>47</td>
<td>34</td>
<td>81</td>
<td>0.65</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>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>46</td>
<td>29</td>
<td>75</td>
<td>0.69</td>
<td>7.31</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>34</td>
<td>51</td>
<td>85</td>
<td>0.69</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>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>47</td>
<td>38</td>
<td>85</td>
<td>0.67</td>
<td>2.03</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>42</td>
<td>75</td>
<td>0.73</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>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>30</td>
<td>40</td>
<td>70</td>
<td>0.82</td>
<td>2.54</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>50</td>
<td>40</td>
<td>90</td>
<td>0.41</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>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>34</td>
<td>43</td>
<td>77</td>
<td>0.45</td>
<td>2.03</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>46</td>
<td>37</td>
<td>83</td>
<td>0.78</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>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>35</td>
<td>42</td>
<td>77</td>
<td>0.60</td>
<td>1.23</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>45</td>
<td>38</td>
<td>83</td>
<td>0.74</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>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>47</td>
<td>29</td>
<td>76</td>
<td>0.66</td>
<td>8.12</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>51</td>
<td>84</td>
<td>0.70</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>
</tr>
</tbody>
</table>

We retained the 598 spectra associated with 1, 7, 9, 11, 15, and 20-day old semi-field reared mosquitoes. We mixed them with all 927 wild spectra and performed the analysis as in the first approach. Figure 4.2, panel E represents the
silhouette coefficients of objects in each cluster showing the quality of clusters was not compromised with the removal of 3, 5, and 25-day old semi-field reared mosquitoes in the analysis. Figure 4.2, panel F and Table 4.2, row $C_k$ represent the results showing no significant difference between spectra collected from semi-field reared and wild mosquitoes of the same species ($p = 0.26$).

The results from $k$-means approach two (Figure 4.2, panel D and F, Table 4.2, rows $B_k$ and $C_k$ and Table 4.3) strongly suggest that the results from $k$-means approach one (Figure 4.2, panel B, and Table 4.2, row $A_k$) were influenced by mosquito age differences and may not be influenced by their source (semi-field or wild). We did not use age classification labels from ovary dissection to control age of wild mosquitoes because the ovary dissection method only determines the physiological age of mosquitoes and cannot infer mosquito age in days [95]. The method classifies mosquitoes as relatively young (not laid eggs) or old (laid eggs) based on egg laying status. This classification can be misleading, as mosquitoes lay eggs after getting blood for egg development. Therefore, a mosquito can be old without a gonotrophic history or young and have laid eggs.

**$K$-means approach three:** We performed a partial least square regression (PLSR) on the spectra to reduce spectra features from 1851 absorbances to ten components and repeated the $k$-means cluster analysis as in the first approach. Feature reduction using PLSR helps reduce noise in data without losing important
information. PLSR reduces features by finding components associated with all features (absorbances) while considering dependent variables (semi-field or wild in our case) [112, 118]. We found no substantial clusters with SC below 0.25 (Figure 4.5, panels A and B) strengthening the results we obtained when age of mosquitoes was controlled, where we found no difference between spectra collected from semi-field reared and wild mosquitoes of the same species. The results further suggest that clustering in the first approach was influenced by age.

Finding from the second and third approaches of \(k\)-means clustering that there is no difference in spectra collected from semi-field reared and wild mosquitoes, while approach one suggests there is a difference, we repeated cluster analysis using hierarchical clustering on the datasets with and without age of mosquitoes controlled.

We applied agglomerative hierarchical clustering, first on a dataset with the age of mosquitoes not controlled (all 823 semi-field reared mosquitoes and 927 wild mosquitoes). When generating a tree, we restricted the number of leaf nodes to thirty for both simplicity of viewing the tree and analysis of how mosquitoes distribute from higher to lower level clusters. Figure 4.6, panel A and B (also Table 4.2, row \(A_h\)), respectively, present the generated hierarchical tree and the bar graph showing formed clusters with more semi-field reared mosquitoes in cluster one and
Figure 4.5: **Two-dimensional plot of clusters using the first and second PLS components** (Panel A), and **box plots, showing the silhouette coefficient of each spectrum (object) in its associated cluster** (Panel B) when partial least squares was applied to reduce the data dimension before clustering more wild mosquitoes in cluster two. The chi-square test found the difference to be significant ($p < 0.01$), which agrees with the results of $k$-means approach one.

Table 4.4, column A presents the distributions of semi-field and wild mosquitoes in each of the thirty nodes showing almost all nodes containing both types of mosquitoes. Having both semi-field and wild mosquitoes in most of the
formed clusters (node) at the level of thirty clusters (nodes) strongly suggest that
the source of mosquitoes was not the criterion used in forming two clusters.

Figure 4.6: Hierarchical tree and bar graphs showing distributions of semi-field reared and wild mosquitoes in clusters formed by hierarchical cluster analysis: In panels A and B, the age of mosquitoes was not controlled \( (p < 0.01) \); in C and D, the age structure of semi-field reared mosquitoes was fit to an exponential decay distribution to match the published age structure of wild mosquitoes \( (p = 0.76) \); and in E and F, semi-field reared mosquitoes at 3, 5, and 25-days old were omitted from the analysis \( (p = 0.13) \).
Table 4.4: **Number and type of mosquitoes in leaf nodes of the hierarchical tree:** A) Age of mosquitoes not controlled; B) Age of mosquitoes controlled by selecting age of laboratory-reared mosquitoes to fit the published age distribution of wild mosquitoes; C) Laboratory-reared mosquitoes at age 3, 5, and 25-day old excluded in the analysis.

<table>
<thead>
<tr>
<th>Leaf node</th>
<th>No. semi-field mosquitoes</th>
<th>No. wild mosquitoes</th>
<th>No. semi-field mosquitoes</th>
<th>No. wild mosquitoes</th>
<th>No. semi-field mosquitoes</th>
<th>No. wild mosquitoes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93</td>
<td>42</td>
<td>3</td>
<td>5</td>
<td>114</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>92</td>
<td>151</td>
<td>1</td>
<td>5</td>
<td>64</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>10</td>
<td>34</td>
<td>18</td>
<td>90</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>89</td>
<td>11</td>
<td>23</td>
<td>28</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>32</td>
<td>26</td>
<td>17</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>158</td>
<td>69</td>
<td>20</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>5</td>
<td>15</td>
<td>28</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>23</td>
<td>39</td>
<td>26</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>101</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>27</td>
<td>78</td>
<td>18</td>
<td>11</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>12</td>
<td>15</td>
<td>25</td>
<td>55</td>
<td>43</td>
</tr>
<tr>
<td>12</td>
<td>110</td>
<td>68</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>38</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>21</td>
<td>8</td>
<td>10</td>
<td>82</td>
<td>194</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>49</td>
<td>71</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>24</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>6</td>
<td>23</td>
<td>31</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>24</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>24</td>
<td>48</td>
<td>32</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>32</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>111</td>
<td>73</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

We repeated hierarchical clustering on the datasets with the age of mosquitoes controlled, first, by an exponential decay curve. We found a hierarchical tree (Figure 4.6 panel C) with no significant difference ($p = 0.76$) in the distribution
of semi-field reared and wild mosquitoes between the two-formed clusters (Figure 4.6, panel D and Table 4.2, row $B_h$). Table 4.4, column B represents mosquito distributions in each of the thirty nodes. Most of the nodes consist of both semi-field reared and wild mosquitoes, further suggesting that clustering is independent of the source of mosquitoes.

Second, we repeated hierarchical clustering by removing spectra associated with 3, 5, and 25-day old semi-field reared mosquitoes from the data set. Figure 4.6, panel E represents a hierarchical tree showing no significant difference ($p = 0.13$) in the distribution of semi-field reared and wild mosquitoes between the clusters (Figure 4.6, panel F and Table 4.2, row $C_h$). Table 4.4, column C presents mosquito distributions in each of the thirty nodes, showing the same trend of each node consisting both semi-field reared and wild mosquitoes.

These results agree with the results obtained in $k$-means cluster analysis, strongly suggesting that there is no difference between spectra collected from semi-field reared and wild mosquitoes of the same age and species.

4.5 Discussion

In this study, we investigated whether there is any significant difference between NIR spectra collected from semi-field reared and spectra collected from wild
mosquitoes. Our results show that $k$-means cluster analysis on the mixture of spectra without controlling the age of semi-field reared mosquitoes produced clusters apparently associated with the source of the spectra. This could suggest that there is a difference between spectra collected from semi-field reared mosquitoes and those collected from the wild. However, different factors apart from the source of the spectra may have contributed to the results. The age of a mosquito is one of the most important factors to consider, as different studies [75, 120, 128] have 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 also can influence the cluster formation. Ntamatungiro et al. [95] showed there is an influence of physiological status of a mosquito on the spectra.

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 mosquitoes. The results showed no influence of the source of mosquitoes on forming clusters. This means that in the first approach, age probably played an important role in cluster formation. 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 semi-field reared mosquitoes, results showed no influence on cluster formation.

Since partial least squares analysis has been shown to be effective for age-classification of semi-field reared mosquitoes, we performed partial least square analysis on the spectra to reduce the number of features before we did cluster analysis. Feature reduction using PLS can help during analysis by reducing noise in data without losing important information. Initially, the spectra had 1851 features, which can introduce errors during cluster analysis. PLS discards only a little information when reducing features; instead it finds components associated with all features while considering dependent variables [112, 118]. When we applied PLS and performed \( k \)-means clustering on the reduced features (ten components), we found very poor clustering, with average SCs below 0.21, which indicates that there is no clustering tendency in the data [114, 138]. These results strengthened the results obtained when the age of semi-field reared mosquitoes was controlled.

We finally applied hierarchical clustering to test the reproducibility of the results from \( k \)-means clustering. The distribution of semi-field reared and wild mosquitoes in the formed clusters was not significant when age of mosquitoes was controlled (\( p = 0.13 \)), suggesting there is no difference between spectra collected from semi-field reared and wild mosquitoes of the same age and species. Having two clustering methods with different clustering mechanisms reaching the same
conclusion, we fail to reject the null hypothesis that there is no significant difference between the spectra collected from semi-field reared and those from wild mosquitoes of the same age and species.

4.6 Conclusion - No difference between spectra from semi-field and wild mosquitoes

Our study concludes that there is no difference between spectra collected from semi-field reared and wild mosquitoes of the same age and species. While further studies may be required to explore a more appropriate way to estimate age of wild mosquitoes, these results strengthen the ongoing practice of training models to estimate age of wild mosquitoes using spectra collected from semi-field reared mosquitoes [62, 120]. The reliability of the age estimates from the model might still be questioned, which is acceptable, as model estimates are not always expected to be accurate [105, 134, 146]. Despite of this known caveat, the most important advantage of using models is to give insight into situations where it is difficult to get the truth [28, 105]. Getting actual age in days of wild mosquitoes is difficult, tedious, time inefficient, and expensive. Therefore, the practice of applying a model trained on first generation semi-field reared mosquitoes to estimate wild mosquitoes might not be ideal, but the results from this study show that it might be reliable enough to give insight into age structure of a wild mosquito population, especially
when complemented with other existing knowledge on age structure of wild mosquitoes.

4.7 Transition to Chapter 5

Based on results in Chapters 2 and 4 (also in [84, 83]), in Chapter 5 we train ANN models on semi-field raised *An. arabiensis* to classify wild *An. arabiensis* into less than or greater or equal to seven days old.
CHAPTER 5

An Artificial Neural Network Model to Estimate Age Class in Days of Wild *An. arabiensis*

This chapter [86] will be available in bioRxiv as a pre-print and submitted to the *American Journal of Tropical Medicine and Hygiene* for publication consideration as an original paper.

Abstract

**Introduction:** The age of wild mosquitoes is estimated by dissecting their ovaries to determine whether they have laid eggs. Mosquitoes that have laid eggs are assumed to be older than those which have not laid eggs. This assumption is not always true, as mosquitoes lay eggs after accessing blood for egg development. Hence, a mosquito can be old without egg laying history or can be young and have laid eggs. Having a model which estimates the age of wild mosquitoes in days can complement this interpretation, providing more reliable age estimates. A number of studies have demonstrated that machine learning techniques applied to near-infrared spectra can estimate the age, species, and infectious state of laboratory and semi-field raised mosquitoes with accuracies between 80 - 99%. To train an NIRS model that estimates age in days of wild mosquitoes requires samples
of wild mosquitoes with age labels in days. Such labels are difficult and expensive to get. However, a recent study shows no clear difference between near-infrared spectra (NIRS) collected from semi-field reared and wild mosquitoes of the same species.

**Methods and results:** Hence, we trained ANN binary classifiers on near-infrared spectra collected from semi-field raised *An. arabiensis* (*N* = 870) and applied it to classify wild collected *An. arabiensis* (*N* = 912). Lacking chronological ages of wild *An. arabiensis*, in addition to reporting the accuracy of the classifier on the out-of-sample semi-field *An. arabiensis*, we score the accuracy of the model using parity status determined by dissection of wild *An. arabiensis* with an assumption that most parous mosquitoes should be classified as greater or equal to seven days old. The model scored an accuracy of 98.3% and 74.7% when applied on out-of-sample semi-field *An. arabiensis* and wild *An. arabiensis*, respectively, to classify mosquitoes into either less than or greater or equal to seven days old.

Knowing a wild mosquito may have laid eggs at the age less than seven days old (at least at five days old), we retrained a binary classifier to classify wild *An. arabiensis* into either less than or greater than or equal to five days old, and the classifier scored an accuracy of 73.2%.

**Conclusion:** Therefore, for more reliable age estimates, we recommend complementing age estimates from ovary dissection with age estimates from an ANN binary classifier. Since NIRS is a high-throughput, non-invasive, reagent-free,
and simple to use technique, complimenting it with ovary dissection provides age estimates from a more statistically acceptable sample size.

5.1 Introduction to mosquito age estimates from ovary dissection

Before female mosquitoes lay eggs, they require a blood meal for egg development [74, 89]. In the process of acquiring blood, they may access pathogen-infected blood, rendering them potentially infectious [10]. Therefore, knowing their parity status is useful information in the fight against infectious diseases transmitted by mosquitoes [4, 11, 30, 32]. Knowledge of parity status of mosquitoes is applied mainly: i) to determine the vectorial capacity [44, 49] between mosquito populations, where by population with a higher proportion of parous mosquitoes is assumed to be more potentially infectious than a population with a lower proportion of parous mosquitoes [87], and ii) to evaluate the effectiveness of vector control interventions such as long lasting insecticide-treated nets (LLINs) and indoor residual spraying (IRS) [97, 104]. The areas with working interventions are expected to have lower proportions of parous mosquitoes compared to areas with failing or without interventions (with an assumption that mosquitoes, especially anthropophilic mosquitoes [135] will not access blood, hence will not lay eggs).

Parity status also is useful when estimating the age of mosquitoes [4, 14, 22, 30, 31, 107, 32]. Parous mosquitoes are assumed to be older than
nulliparous mosquitoes. This interpretation does not always apply, as mosquitoes lay eggs after accessing blood for egg development [39, 74, 89]. Hence, a mosquito can be old without having a gonotrophic history or young and have laid eggs. Having a model that estimates the age of wild mosquitoes in days can complement this interpretation, giving more accurate estimates.

Previously, models trained on near-infrared (NIR) spectra collected from laboratory and semi-field raised mosquitoes estimated their age in days, species, and infectious state with an average accuracy of 90% [35, 62, 73, 76, 77, 83, 95, 119, 121, 125, 126, 127, 128]. In this study, we train artificial neural network models that estimate the age in days of wild mosquitoes from NIR spectra. Most of the previous studies trained NIR models using partial least square (PLS) as the model architecture. Results published in [83] show that NIR models trained using artificial neural networks (ANN) perform significantly better than NIR models trained using PLS (P < 0.001). Therefore, we trained our NIR model using ANN architectures.

Normally, training models to estimate the actual age of wild mosquitoes requires samples of wild mosquitoes with known age in days. While it is difficult to get samples of wild mosquitoes with known age in days, there is no clear difference between NIR spectra collected from semi-field reared and wild mosquitoes of the same species [84]. Knowing there is no clear difference between spectra collected
from semi-field raised and wild mosquitoes of the same species, Krajacich et al. trained models on lab-reared and semi-field raised \textit{An. gambiae} to estimate age class in days of wild \textit{An. gambiae} [62]. In this study, we train models on semi-field raised \textit{An. arabiensis} to classify wild \textit{An. arabiensis} as less than or greater than or equal to seven days old. Using parity status as a truth (with an assumption that most parous mosquitoes should be classified as greater than or equal to seven days and most nulliparous mosquitoes as less than seven days old) our model classified wild \textit{An. arabiensis} into either less than seven days or greater than seven days old with an accuracy of 74.7%. Since there is a possibility for a mosquito to lay eggs at the age less than seven days (at least five days old) [74, 89], we further trained another model that classifies wild \textit{An. arabiensis} into less than or greater than or equal to five days old. The model scored an accuracy of 73.2%. These results suggest that ANN binary classifier trained on NIR spectra can compliment ovary dissection. While ovary dissection provides physiological age, the ANN model can provide age estimates in days.

5.2 Material and method

In this section, we first provide an ethical clearance used to collect data involved humans. We further state the materials used in the study and describe the modeling process.
**Ethics approval:** Semi-field raised mosquitoes were fed human blood under Ethical Clearance No. IHRDC/EC4/CL.N96/2004, and wild mosquitoes were collected from people’s homes under Ethical Clearance No. IHI-IRB/No 17–2015 provided by the Ifakara Health Institute (IHI) Review Board. Each volunteer (at least eighteen years old) provided an oral consent before becoming involved in the study and was given the right to refuse to participate or to withdraw from the study at any time.

**Mosquitoes:** We used 870 (at least 80 mosquitoes at 1, 3, 5, 7, 9, 11, 15, 20, and 25 days post emergence) semi-field *An. arabiensis* raised under surrounding temperature and light-dark cycles in semi-field systems [41, 93] owned by Ifakara Health Institute (IHI), located at Ifakara in southeastern Tanzania. Mosquitoes were fed 10% sugar solution daily and a human blood meal twice a week (Ethical Clearance No. IHRDC/EC4/CL.N96/2004). More details on how semi-field mosquitoes are raised in this insectary are given in [84].

We also used 912 wild *An. arabiensis* collected from March to October in people’s homes (Ethical Clearance No. IHI-IRB/No 17–2015) using CDC light traps in Minepa, a village located in the Kilombero district of southeastern Tanzania. Mosquitoes that were morphologically identified as members of *An. gambiae* complex were sorted and sent to the Ifakara Health Institute laboratory for polymerase chain reaction (PCR) analysis to identify *An. arabiensis* and for ovary dissection to determine their parity status (egg laying status).
**Spectra collection:** Before scanning, mosquitoes were killed by freezing for 20 minutes and left to equilibrate to room temperature for 30 minutes. We collected spectra from both semi-field and wild *An. arabiensis* using a LabSpec 5000 NIR spectrometer (ASD Inc., Malvern, UK).

**Spectra pre-processing:** We pre-processed spectra from both semi-field raised and wild collected *An. arabiensis* according to the previously published protocol [77]. We then applied $k$-means and hierarchical clustering to determine if there is any significant difference between these two spectra groups (i.e., spectra collected from semi-field raised and wild collected *An. arabiensis*). From both $k$-means and hierarchical clustering, we found no significant difference between these spectra ($p = 0.25$ and $p = 0.13$ for $k$-means and hierarchical clustering, respectively). More results and detailed information on how clustering was applied on these spectra is in Chapter Four and published in [84].

Finding that there is no significant difference between spectra collected from semi-field raised and wild collected *An. arabiensis*, we trained an ANN model on semi-field *An. arabiensis* to classify age of wild collected *An. arabiensis*.

**Model training:** We divided pre-processed spectra from semi-field *An. arabiensis* into either less than or greater or equal to seven days, and label them 0 and 1, respectively. We then merged, randomized, and divided re-labeled spectra into training (75% of 870 spectra) and test (25% of 870 spectra) sets. We used ten
Monte-Carlo cross validations [83, 148] and Levenberg-Marquardt optimization to train an ANN model with one hidden layer, ten neurons, and logistic regression as a transfer function to classify wild collected An. arabiensis into either less than seven days or greater or equal to seven days old. We refer this model as model-seven.

When interpreting the model output, mosquitoes estimated as < 0.5 were considered as < 7 days old and ≥ 0.5 as ≥ 7 days old.

5.3 Results and discussion

Table 5.1 represents the performance of the model-seven when it was applied on out-of-sample test sets (25% out-of-sample semi-field raised An. arabiensis), showing an accuracy of 98.3%. More results on the performance of this model on the test samples and independent test sets are in [83]. The results presented in Table 5.1, in Chapter 2, and in [83] show model-seven was not overfit on the training dataset.

Following the performance of model-seven on test samples of semi-field raised An. arabiensis presented in Table 5.1, in Chapter 2, and in [83], we applied model-seven to estimate the age class of wild caught An. arabiensis whose parity status are known. Model-seven classified 72% (N = 912) of wild An. arabiensis as less than seven days old and 28% (N = 912) as greater or equal to seven days old. Lacking chronological age labels of wild An. arabiensis, we cannot validate directly the accuracy of the model on estimating age in days of wild mosquitoes.
Table 5.1: Performance of ANN binary classifiers on estimating the age class in days of 25% out-of-sample semi-field raised *An. arabiensis*. TP: Number of semi-field *An. arabiensis* correctly classified as $\geq 7$ days old; TN: Number of semi-field *An. arabiensis* correctly classified as $< 7$ days old; P: Total number of semi-field *An. arabiensis* that are $\geq 7$ days old; N: Total number of semi-field *An. arabiensis* that are $< 7$ days old.

| Metric | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|--------|--------------|----------------|----------------|               |
| Score  | 98.3 $\pm$ 0.9 | 98.5 $\pm$ 1.3 | 98.3 $\pm$ 0.6 |

Instead, based on findings in Chapter 4 and in [84], the accuracy of model-seven is assumed to be equivalent to the accuracy presented in Table 5.1. We also analyzed the distribution of nulliparous and parous mosquitoes in each estimated class. The assumption is that most nulliparous mosquitoes should be classified as less than seven days old, and most parous mosquitoes as seven days or more [74, 89]. Using parity status as the truth and Equations 5.1, 5.2, 5.3, and 5.4, we calculated the accuracy, sensitivity (recall) (the ability of the model to classify *An. arabiensis* correctly which are greater or equal to seven days old), specificity (the ability of the model to classify *An. arabiensis* correctly that are less than seven days old), and precision (proportion of wild *An. arabiensis* correctly classified as greater than or equal to seven days out of all wild *An. arabiensis* classified by the model-seven as greater than or equal to seven days old), respectively.
Let

- TP = Number of wild *An. arabiensis* correctly classified as ≥ 7 days old,
- FP = Number of wild *An. arabiensis* wrongly classified as ≥ 7 days old,
- TN = Number of wild *An. arabiensis* correctly classified as < 7 days old,
- P = Number of wild *An. arabiensis* that are parous (assumed to be ≥ 7d), and
- N = Number of wild *An. arabiensis* that are nulliparous (assumed to be < 7d).

Then

\[
\text{Sensitivity of the model} = \frac{TP}{P}, \quad (5.1)
\]

\[
\text{Specificity of the model} = \frac{TN}{N}, \quad (5.2)
\]

\[
\text{Accuracy of the model} = \frac{TP + TN}{P + N}, \quad (5.3)
\]

\[
\text{Precision of the model} = \frac{TP}{TP + FP}. \quad (5.4)
\]

Tables 5.2 - 5.3 and Figure 5.1 present the results showing the performance of the model-seven with an accuracy of 74.7%.
Table 5.2: Performance of the ANN binary classifier (model-seven) trained on semi-field raised *An. arabiensis* (N = 653) on classifying wild *An. arabiensis* (N = 912) into less than or ≥ 7 days old.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score (%)</td>
<td>74.7</td>
<td>53.8</td>
<td>82.9</td>
<td>55.8</td>
</tr>
</tbody>
</table>

Table 5.3: The number of parous and nulliparous wild *An. arabiensis* in each age class estimated by model-seven after the model was applied to classify wild *An. arabiensis* into either less than or ≥ 7 days old.

<table>
<thead>
<tr>
<th>Parity Status</th>
<th>Age group in days</th>
<th>Nulliparous (N)</th>
<th>Parous (N)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates</td>
<td>&lt; 7 days</td>
<td>82% (N = 541)</td>
<td>18% (N = 120)</td>
<td>661</td>
</tr>
<tr>
<td></td>
<td>≥ 7 days</td>
<td>44% (N = 111)</td>
<td>56% (N = 140)</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>652</td>
<td>260</td>
<td>912</td>
</tr>
</tbody>
</table>

The performances of the model-seven presented in Tables 5.2 - 5.3 and Figure 5.1 should be taken with caution. First, there is a possibility for a mosquito to lay eggs when it is less than seven days old. If a female mosquito mates on the day she emerges and accesses a blood meal the same day or the next day after mating, she can lay eggs when she is five days old [74, 89]. Also, a female mosquito can be more than seven days old without any gonotrophic history, as she needs blood to lay eggs [74, 89]. Second, the process of dissecting mosquito ovaries to score their parity status also can be a source of errors, as a mosquito can be labeled mistakenly as parous, while it is nulliparous; and it can be scored as nulliparous, while it is parous. Therefore, the model’s true accuracy could be higher or less than 74.7%.
To account for wild mosquitoes that might have laid eggs at five days old (if any) we trained another ANN model to classify *An. arabiensis* into less than five days or greater or equal to five days old. We refer to this model as model-five. Table 5.4 and Figure 5.2 represent the results of the model-five when was tested on the 25% out-of-sample semi-field raised *An. arabiensis* showing an average accuracy of 97%. Figure 5.3 and Table 5.5 represent the performance of model-five applied to
classify wild *An. arabiensis*, showing an insignificant change in model performance (accuracy = 73.2%), suggesting few or no mosquitoes laid eggs at five days old.

Table 5.4: **Accuracy of model-five in estimating age class of 25% out-of-sample semi-field raised** *An. arabiensis*. TP: Number of semi-field *An. arabiensis* correctly classified as ≥ 5 days old; TN: Number of semi-field *An. arabiensis* correctly classified as < 5 days old; P: Total number of semi-field *An. arabiensis* that are ≥ 5 days old; N: Total number of semi-field *An. arabiensis* that are < 5 days old.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score (%)</td>
<td>97.0 ± 1.1</td>
<td>97.7 ± 0.5</td>
<td>96.1 ± 1.3</td>
</tr>
<tr>
<td><strong>Formula</strong></td>
<td>(\frac{TP + TN}{P + N})</td>
<td>(\frac{TP}{P})</td>
<td>(\frac{TN}{N})</td>
</tr>
</tbody>
</table>

Table 5.5: **The number of parous and nulliparous wild *An. arabiensis* in each age class estimated by model-five after the model was applied to classify wild *An. arabiensis* into either less than or ≥ 5 days old.**

<table>
<thead>
<tr>
<th>Parity Status</th>
<th>Age group in days</th>
<th>Nulliparous</th>
<th>Parous</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt; 5 days</td>
<td>82% (N = 525)</td>
<td>18% (N = 117)</td>
<td>661</td>
</tr>
<tr>
<td></td>
<td>≥ 5 days</td>
<td>44% (N = 127)</td>
<td>56% (N = 143)</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>652</td>
<td>260</td>
<td>912</td>
</tr>
</tbody>
</table>

In addition, we computed a different metric, the Jaccard similarity coefficient (JC) [54, 94] to determine the similarity of the results from model-seven and ovary dissections when applied separately to classify age of the same sample of wild
Figure 5.2: Receiver operating curve (ROC) with area under the curve (AUC) presenting the performance of model-five when was applied to classify 25% out-of-sample semi-field raised *An. arabiensis* into less than or ≥ 5 days old mosquitoes as illustrated in Figure 5.4. Here, we disregard the mechanism behind each of the two methods in classifying mosquitoes. Our focus is on the similarity of the results (age class estimates) from ovary dissection and from model-seven.
Figure 5.3: Percentage of nulliparous and parous mosquitoes in each age class estimated by model-five after it was applied to classify wild *An. arabiensis* (N = 912) into either less than or ≥ 5 days old

Let

- A = set of age class estimates according to ovary dissection, and
- B = set of age class estimates according to model-seven.

Then

$$JC(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (5.5)$$
Interpretation:

- Jaccard coefficient (JC) ranges from 0 to 1,
- High JC = The sets are similar (JC = 1, sets are same), and
- Low JC = The sets are dissimilar (JC = 0, sets are different).

Using information given in Table 5.3 and Equation 5.5, we found the Jaccard similarity coefficient (JC) between age class estimates from model-seven and Detinova ovary dissection to be 0.75.

While in our case the outputs from Equations 5.3 and 5.5 might be similar, it is the interpretation that is different. A Jaccard similarity coefficient of 0.75 means that there is a 75% chance that Detinova ovary dissection and model-seven will classify a mosquito into the same age class and a 25% chance they will classify a mosquito into different age classes. These results suggest that model-seven can be a reliable method to complement the Detinova ovary dissection method. Since NIRS is a high-throughput technique (hundreds of mosquitoes can be scanned per day, while one hundred dissections would be a good day’s output from an experienced expert) and 75% similar to Detinova dissection, complementing Detinova ovary dissection with model-seven allows drawing conclusions on the age composition of a particular wild mosquito population not only from more accurate estimates but also
Figure 5.4: Using the Jaccard similarity coefficient to determine similarity in outputs from models trained on semi-field raised mosquitoes to estimate the age class of wild mosquitoes and from Detinova ovary dissection from 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.

5.4 Conclusion

This study trained an ANN binary classifier on near-infrared spectra to estimate the age class in days of wild mosquitoes to compliment physiological age estimates from
ovary dissections. Assuming that parous mosquitoes should be classified as greater or equal to seven days old and most nulliparous mosquitoes as less than seven days old, the ANN binary classifier trained on semi-field raised mosquitoes classified wild mosquitoes into either less that seven days or greater or equal to seven days old with an accuracy of 74.7%. Using the Jaccard similarity coefficient, there is a 75% chance for ANN binary classifier and ovary dissection to classify a wild mosquito into the same age class. These scores suggest that an ANN binary classifier can compliment ovary dissections to estimate the age class of a particular population of mosquito vectors of infectious diseases, and provide more reliable estimates, especially since NIRS is a high-throughput technique compared to ovary dissection. While an experienced expert can dissect not more than one hundred mosquitoes in a day, hundreds of mosquitoes per day can be NIRS scanned. Therefore, we recommend an ANN binary classifier trained on semi-field raised mosquitoes as a complementary method to ovary dissections to estimate the age class in days of wild mosquitoes.
CHAPTER 6

Conclusions and Future Work

In each of the previous Chapters, we have stated the specific conclusions drawn. In this Chapter, we provide general conclusions drawn from all four objectives and give recommendations for future studies.

6.1 General conclusions

Malaria is one of the deadly infectious diseases in the world [144]. Just in 2018, 228 million cases of malaria were reported worldwide [144]. According to the 2019 World Health Organization (WHO) report, about 180,000 lives were saved between 2010 (585,000 reported deaths in 2010) and 2018 (405,000 reported deaths in 2018) [144]. This success was achieved through a combination of approaches ranging from proper treatment of the disease, vector management, vaccine, and public education on malaria. Despite the current success in the fight against malaria, new approaches and interventions are needed to save more lives and completely eradicate the disease. In this study, we explored new tools to help manage malaria vectors. We demonstrated the validity of neural networks applied to near-infrared spectra as a tool to either complement or replace existing tools to estimate parity status and age (in days) of wild mosquitoes. Parity status and age (in days) of mosquitoes are
important to monitor and evaluate the effectiveness of the malaria vector interventions in place (i.e., insecticide treated nets (ITNs), indoor residual spraying (IRS), and larviciding).

In Chapter 2, we explored ways to improve the current accuracy of the models trained on near-infrared spectra to estimate age of laboratory-reared and semi-field raised mosquitoes. We find that using artificial neural network (ANN) as an architecture yields models that score higher accuracies than when the same models are trained using a partial least squares (PLS) model architecture. Also, irrespective of the model architecture used, a directly trained binary classifier performs better than a regressor interpreted as a binary classifier.

In Chapter 3, we trained ANN models on autoencoded near-infrared spectra to estimate parity status of wild mosquitoes. The models scored an average accuracy of 93%. The work presented in Chapter 3 is not only the first to train models that estimate parity status of wild mosquitoes but also to apply ANN on autoencoded NIR spectra to train models. The accuracy suggests ANN models trained on autoencoded NIRS as an appropriate alternative to ovary dissection. Since NIRS is a high-throughput technique, complementing ovary dissection with NIRS provides a more statistically acceptable sample size to draw conclusions on the transmission capacity of a certain mosquito population.
The problem of training models that estimate the age of wild mosquitoes is the lack of samples of wild mosquitoes with known age in days to train and test the models. Methods such as mark-release-recapture and molecular techniques are expensive and time consuming. As an alternative, models are trained on semi-field raised mosquitoes to classify wild mosquitoes into either less than or equal or greater than seven days, but no study had justified this practice. Results presented in Chapter 4 are the first to justify the on-going practice of training models on semi-field raised mosquitoes to estimate the age of wild mosquitoes. We applied clustering techniques to determine if there is any significant difference between NIR spectra collected from semi-field raised and spectra collected from wild mosquitoes of the same species. With $k$-means ($p = 0.25$) and hierarchical ($p = 0.13$) clustering, we failed to reject the hypothesis that there is no significant difference between NIR spectra collected from semi-field raised and wild collected An. arabiensis.

In Chapter 5, we trained an ANN model on near-infrared spectra collected from semi-field raised An. arabiensis to classify the ages of wild An. arabiensis whose parity status are known. Using Jaccard similarity coefficient, we found that there is a 75% chance that our ANN model trained on semi-field raised mosquitoes and ovary dissections classify wild mosquitoes into the same age class.
6.2 Future work

We suggest future studies to:

1. Train ANN models on autoencoded NIR spectra to estimate parameters of wild mosquitoes other than parity status, i.e., species, infectious state, and type of blood meal;

2. Apply transfer learning techniques to train a deep neural network model that extrapolates easily to different tasks and spectra with different characteristics than those used to train the model; and

3. Develop a graphical user interface (GUI) that will have models running behind the scene allowing users such as field technicians and public health officials to use the models with little knowledge of the mechanisms behind the models.

We discuss each of these in turn.

6.2.1 Training ANN models on autoencoded NIR spectra to estimate parameters of wild mosquitoes other than parity status

In Chapter Three, we demonstrated the effectiveness of ANN models trained from autoencoded NIR spectra to estimate the parity status of wild mosquitoes. We recommend future studies to repeat the analysis on larger datasets. We also
recommend future studies to train similar models that estimate species, type of
blood meal, and infectious state of wild mosquito vectors of infectious diseases.

6.2.2 Applying transfer learning techniques to train a deep neural
network model

Models trained on near-infrared spectroscopy (NIRS) have successfully
estimated different parameters such as infectious state [62, 73, 127], age, and
species [35, 62, 75, 76, 81, 82, 84, 83, 119, 120, 121, 126, 127] of laboratory and
semi-field reared mosquitoes. All these models were developed on near-infrared
spectra with the same characteristics (i.e., spectra scanned from mosquitoes
collected from same geographical region, either killed or preserved the same way) to
perform single task. Having models which perform a single task or use spectra with
very specific characteristics impair scaling up of this promising technique in the
fight against infectious diseases such as malaria, zika, dengue, and chikungunya.
Since these models were trained on spectra of the same characteristics, they cannot
correctly estimate parameters of mosquitoes using spectra that have different
characteristics from those used to train the model. The practice of training models
to perform a single task can be practical and useful only when there is sufficient
data from each scenario (phenomenon) to train and test that model. However, in
some very important cases, data availability is an issue. For example, if a model to
estimate the infectious state of a mosquito is needed, sufficient spectra to train and
test that model is always challenging. Alternatively, to increase the chances of getting a reasonable number of samples of infected mosquitoes, one could collect infected mosquitoes from different geographical regions, but this is also limited with the current practice of training models using spectra of the same characteristics.

Neural networks are known to learn successfully important information from noisy datasets with diverse characteristics and to extrapolate the learned knowledge to estimate instances which were not presented in the training set [68, 137]. While most machine learning model architectures are designed to perform a single task, deep neural networks (DNN), through a technique known as transfer learning, can be trained to perform multiple related tasks. Transfer learning is a machine learning technique that allows a previously trained model for a particular task to be modified for another related task [16, 141].

Transfer learning commonly is applied in the fields of image and language studies. Google, Microsoft, and Oxford have ongoing projects that apply transfer learning technique on a deep learning model pre-trained on a large database such as ImageNet [65, 96] to develop models (listed below) that can classify any type of image or text irrespective of whether that image or text was represented in the training set:

- Google’s Inception Model [47].
- Google’s Word2vec Model [48, 80].
• Microsoft’s ResNet Model [56].

• Stanford’s GloVe Model [100].

• Oxford’s VGG Model [129].

Therefore, we suggest future studies to train a deep neural network (DNN) model on spectra scanned on mosquitoes collected from different geographical regions, using different scanning instruments, either killed or preserved using different methods, and apply transfer learning technique on a trained DNN to estimate different parameters of mosquitoes such as age, species, and infectious state.

6.2.3 Developing a graphical user interface (GUI)

The idea here is to develop an application or web product with a DNN model trained using a transfer learning technique running behind the scenes. The application or web product should be friendly enough to allow users such as field technicians to use it with little knowledge of the mechanism behind the model. The product should also have options that allow users to choose and estimate desired parameters of mosquitoes.
6.3 Concluding statement

We expect results from this dissertation provide more evidence to support the use of machine learning models trained from near-infrared spectra as one of the methods to estimate required parameters of mosquito vectors of infectious diseases. Using machine learning applied to NIRS to estimate different mosquito parameters is time and cost-effective compared to the current techniques such as ovary dissections, polymerase chain reaction (PCR) (for species identification), and enzyme-linked immunosorbent assay (ELISA) (for infectious state identification). No reagents are required when using NIRS to estimate mosquito parameters. After the initial expenditure for an NIR spectrometer ($58,040 USD) and analysis of about 40,000 samples, machine learning models trained from NIR spectra become more cost-effective than ovary dissection, PCR, and ELISA (at IHI, it costs $1.5 USD per mosquito to run either PCR or ELISA).
BIBLIOGRAPHY


138


