Autoencoders and Artificial Neural Networks Applied to Near-infrared Spectra to Estimate Parity Status of Wild *An. gambiae* s.s and *An. arabiensis*

Masabho P. Milali[1,2]*, Samson S. Kiware[1,2], Naveen Bansal[1], Sedar Bozdag[3], Jacques D. Charlwood[4], Floyd E. Dowell[5], George F. Corliss[1,6], Maggy T. Sikulu-Lord[1,7], Richard J. Povinelli[6]

1. Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, Wisconsin, United States of America
2. Ifakara Health Institute, Environmental Health and Ecological Sciences Thematic Group, Ifakara, Tanzania
3. Department of Computer Science, Marquette University, Milwaukee, Wisconsin, United States of America
4. Liverpool School of Tropical Medicine, Liverpool, United Kingdom
5. USDA, Agricultural Research Service, Center for Grain and Animal Health Research, 1515 College Avenue, Manhattan KS 66502
6. Department of Electrical and Computer Engineering, Marquette University, Milwaukee, Wisconsin, United States of America
7. Queensland Alliance of Agriculture and Food Innovation, The University of Queensland, Brisbane, Queensland, Australia

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 are currently used to determine 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 age in days and 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 parity status of wild mosquitoes. 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; *Anopheles arabiensis* collected from Minepa, Tanzania (Minepa-ARA); *Anopheles gambiae s.s* collected from Muleba, Tanzania (Muleba-GA); *Anopheles gambiae s.s* collected from Burkina Faso (Burkina-GA); and *An.gambiae s.s* from Muleba and Burkina Faso combined (Muleba-Burkina-GA). We train ANN models on datasets with spectra only pre-processed 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 ± 2.8 $(\mathrm{N} = 912)$, 68.7 ± 4.8 $(\mathrm{N} = 140)$, 80.3 ± 2.0 $(\mathrm{N} = 158)$, and 75.7 ± 2.5 $(\mathrm{N} = 298)$, respectively. With the autoencoder, ANN models tested on out-of-sample data scored $97.1 \pm 2.2\%, (\mathrm{N} = 912)$, $89.8 \pm 1.7\%$ $(\mathrm{N} = 140)$, $93.3 \pm 1.2\%$ $(\mathrm{N} = 158)$, and $92.7 \pm 1.8\%$ $(\mathrm{N} = 298)$ accuracies for Minepa-ARA, Muleba-GA, Burkina-GA, and Muleba-Burkina-GA, respectively. These results show that a combination of an autoencoder and ANNs on NIR spectra 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.