METHOD ARTICLE

Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 1; peer review: 1 approved with reservations, 1 not approved]

Mario González Jiménez, Simon A. Babayan, Pegah Khazaeli, Margaret Doyle, Finlay Walton, Elliott Reedy, Thomas Glew, Mafalda Viana, Lisa Ranford-Cartwright, Abdoulaye Niang, Doreen J. Siria, Fredros O. Okumu, Abdoulaye Diabaté, Heather M. Ferguson, Francesco Baldini, Klaas Wynne

1School of Chemistry, University of Glasgow, Glasgow, G12 8QQ, UK
2Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, UK
3Department of Medical Biology and Public Health, Institut de Recherche en Science de la Santé (IRSS), Bobo-Dioulasso, Burkina Faso
4Environmental Health & Ecological Sciences Department, Ifakara Health Institute, Off Mlabani Passage, PO Box 53, Ifakara, Tanzania

* Equal contributors

Abstract

Despite the global efforts made in the fight against malaria, the disease is resurging. One of the main causes is the resistance that Anopheles mosquitoes, vectors of the disease, have developed to insecticides. Anopheles must survive for at least 10 days to possibly transmit malaria. Therefore, to evaluate and improve malaria vector control interventions, it is imperative to monitor and accurately estimate the age distribution of mosquito populations as well as their population sizes. Here, we demonstrate a machine-learning based approach that uses mid-infrared spectra of mosquitoes to characterise simultaneously both age and species identity of females of the African malaria vector species Anopheles gambiae and An. arabiensis. mid-infrared spectroscopy-based prediction of mosquito age structures was statistically indistinguishable from true modelled distributions. The accuracy of classifying mosquitoes by species was 82.6%. The method has a negligible cost per mosquito, does not require highly trained personnel, is rapid, and so can be easily applied in both laboratory and field settings. Our results indicate this method is a promising alternative to current mosquito species and age-grading approaches, with further improvements to accuracy and expansion for use with other mosquito vectors possible through collection of larger mid-infrared spectroscopy data sets.
Keywords
Malaria, Anopheles gambiae, Anopheles arabiensis, Vector control, Machine learning, Mid-infrared spectroscopy

Corresponding authors: Francesco Baldini (Francesco.Baldini@glasgow.ac.uk), Klaas Wynne (klaas.wynne@glasgow.ac.uk)

Author roles: González Jiménez M: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Babayan SA: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Khazaell P: Investigation, Writing – Review & Editing; Doyle M: Investigation, Writing – Review & Editing; Walton F: Investigation, Writing – Review & Editing; Reedy E: Investigation, Writing – Review & Editing; Glew T: Investigation, Writing – Review & Editing; Viana M: Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Ranford-Cartwright L: Conceptualization, Resources, Writing – Review & Editing; Niang A: Conceptualization, Writing – Review & Editing; Siria DJ: Conceptualization, Writing – Review & Editing; Okumu FO: Conceptualization, Funding Acquisition, Writing – Review & Editing; Diabaté A: Conceptualization, Writing – Review & Editing; Ferguson HM: Conceptualization, Funding Acquisition, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; Baldini F: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; Wynne K: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust through a Intermediate Fellowship in Public Health and Tropical Medicine to FO [102350]. This work was also supported by The Engineering and Physical Sciences Research Council (EPSRC) [EP/J009733/1, EP/K034995/1, EP/N508792/1, and EP/N007417/1] and Medical Research Council (MRC) [MR/P025501/1]. FB is supported by an AXA Research Fund fellowship [14-AXA-PDOC-130] and a European Molecular Biology Organization (EMBO) Long Term fellowship [43-2014]. MV is funded under the MRC/Department for International Development Concor-dat agreement, which is part of EU EDCTP2 programme [MR/N015320/1]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 González Jiménez M et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: González Jiménez M, Babayan SA, Khazaell P et al. Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 1; peer review: 1 approved with reservations, 1 not approved] Wellcome Open Research 2019, 4:76 (https://doi.org/10.12688/wellcomeopenres.15201.1)

First published: 01 May 2019, 4:76 (https://doi.org/10.12688/wellcomeopenres.15201.1)
Between 2000 and 2015, insecticide-based control interventions targeting mosquito vectors averted an estimated 537 million malaria cases\(^1\). Nevertheless, malaria still kills hundreds of thousands of people each year (445,000 in 2016), mainly in sub-Saharan Africa\(^2\). Additionally, there is concern that progress may have stalled after more than a decade of success in global malaria control\(^3\). Of major concern is the increase in insecticide resistance among mosquito populations throughout Africa\(^1\), which is degrading the lethality and effectiveness of vector control tools, notably indoor residual spraying (IRS) and long-lasting insecticide treated nets (LLINs) which have been the cornerstones of malaria control in the past decades\(^4\). Indeed, much of the effectiveness of LLINs and IRS comes from community-wide reductions in vector population size, not merely from preventing people from getting bitten\(^5\).

Measurement of female mosquito vector survival is an important biological determinant of malaria transmission intensity\(^6\). This is because malaria parasites (Plasmodium spp.) require more than 10 days of incubation inside female mosquito vectors (extrinsic incubation period, EIP) before they become infectious\(^7\). While there is uncertainty about mosquito survival in the field, crude estimates suggest the median lifespan of African malaria vectors is 7–10 days\(^7\). Thus, only relatively old mosquitoes can transmit the parasite\(^8\). As a result, even minor reductions in mosquito survival can have exponential impacts on pathogen transmission\(^9\). Consequently, accurate and high-resolution estimation of both mosquito abundance and longevity is essential for the assessment of the impact of these and other control measures.

Despite the crucial importance of mosquito demography to vector control, there are few reliable tools for rapid, high-throughput monitoring of mosquito survival in the wild. Conventionally, mosquito age has been approximated by classifying females (the only sex that transmits malaria) into groups based on their reproductive status as assessed through observation of their ovarian tracheoles\(^10\). This widely-employed technique distinguishes females who have not yet laid eggs (nulliparous) from those that have laid at least one egg batch (parous), with the latter group assumed to be older than the former because the gonotrophic cycle between blood feeding and oviposition takes ~4 days. While useful for approximating general patterns of survival\(^11\), this method is crude and cannot distinguish between females who have laid eggs only once or multiple times. Alternatively, more refined methods have been developed to estimate the number of gonotrophic cycles a female mosquito has gone through based on follicular relics or dilatations formed during each oviposition\(^12\), although the conversion between gonotrophic cycles and actual age is imprecise (especially now that LLINs arelimiting regular access to blood-meals)\(^13\). While an improvement on the simple parity classification method, this approach is extremely technically demanding and time-consuming\(^4\). Additionally, it is unsuitable for analysis of the large sample sizes necessary for estimating mosquito population structure\(^13\).

Given these problems with ovary-based assessment, there has been significant investigation of alternative, molecular-based approaches to estimate mosquito age. These methods include: counting cuticle rings representing daily growth layers of the mosquito skeletal apodemes\(^14\), chromatographic analysis of cuticular hydrocarbon chains\(^15\), assessment of pteridines using fluorescence techniques\(^16\), transcriptomic profiling\(^17\), and mass spectrometric analysis of mosquito protein expression\(^18\). However, even the most accurate, high cost, and/or need of highly trained users suggest that they might not be suitable for application in the field.

In addition to age, identification of mosquito species is crucial for estimation of malaria transmission dynamics. In Africa, the bulk of malaria transmission is carried out by members of the Anopheles gambiae sensu latu and Anopheles funestus sensu latu species complexes\(^19\). The An. gambiae s.l. complex includes several cryptic species that can only be distinguished by molecular analysis\(^20\). Despite being morphologically identical, members of this group vary significantly in behaviour, transmission potential, and response to vector control measurements\(^20\). For example, two major vectors in the An. gambiae s.l. group, An. arabiensis and An. gambiae, can differ in their propensity to enter and rest in houses, their host species choice, breeding conditions, resistance to insecticides, and tolerance to dry climates\(^21\). Currently, An. gambiae s.l. species are best distinguished by polymerase chain reaction (PCR) methods\(^22\), which are time-consuming and expensive, and can thus only be carried out on a subsample of mosquitoes collected during entomological surveillance. Alternative techniques have been developed such as isoenzyme electrophoresis\(^23\) or chromatography of cuticular components\(^24\), but these are also very laborious and have weak discriminatory power\(^25\).

As in the case of age determination, non-PCR-based methods often rely on structural and chemical differences in the cuticle between species. In particular, near-infrared spectroscopy (NIRS) has been evaluated as a general strategy for the discrimination of insects according to their species and other traits since it does not require reagents and holds promise as a fast, practical, non-destructive, and cost-effective method for entomological surveillance. The results obtained to date have proved that the chemical composition of mosquitoes and other insects not only changes between species\(^27–31\), also across different age\(^32\), and during different conditions, resistance to insecticides, and tolerance to dry climates\(^33\). Currently, An. gambiae s.l. species are best distinguished by polymerase chain reaction (PCR) methods\(^33\), which are time-consuming and expensive, and can thus only be carried out on a subsample of mosquitoes collected during entomological surveillance. Alternative techniques have been developed such as isoenzyme electrophoresis\(^35\) or chromatography of cuticular components\(^36\), but these are also very laborious and have weak discriminatory power\(^37\).
is placed in the spectrometer. In addition, the results are normally analysed using Partial Least Squares (PLS) regression, which is prone to over-fitting (i.e. the production of a model that corresponds too closely to a particular set of data and may therefore fail to predict future observations reliably)\cite{9}. This problem commonly arises when the number of samples is relatively small and the number of variables is large.

Here we tested if these limitations can be overcome by shifting the measurement range (25,000–4,000 cm\(^{-1}\)) to the mid-infrared region (4,000–400 cm\(^{-1}\)), employing an attenuated total reflectance (ATR) device to assess the mosquitoes, and modelling the results with supervised machine learning. The mid-infrared absorption spectrum of a mosquito contains a set of discrete well-delineated bands that depend on the fundamental vibrations of the molecules present in the cuticle, providing a wealth of information not present in the near-infrared range, where it is not possible to capture the contributions of different biochemical components of the mosquito to the spectrum and their variations among mosquitoes with different attributes, as shown in \textit{Aedes aegypti} and the diptera \textit{Culicoides sonorensis}\cite{9,50}. However, since the mid-infrared spectral bands are affected in non-trivial ways by the development of a mosquito and the changing composition of the cuticle, it is not possible to predict traits by simply monitoring changes in band intensities\cite{9}.

Here, we show that the use of supervised machine learning\cite{11} allows the determination of the age and species of two major malaria vectors, \textit{An. arabiensis} and \textit{An. gambiae}, from the information contained in their mid-infrared spectra. This is possible because machine learning, unlike standard statistical approaches, can recognise the complex relationships in these traits (mosquito species and mosquito age) and disentangle them from other irrelevant variation\cite{13,14,15,16}. Using this approach, we are able to reconstruct simulated age distributions of mosquito populations with unprecedented reliability. The technique we propose here is time efficient (an analysis takes less than one minute per mosquito), economical, and requires neither reagents nor highly trained operators. It also represents a novel approach to the analysis of insects using spectroscopic techniques, solving some previous drawbacks, and accelerating progress towards reconstructing simulated age distributions of mosquito populations with multiple gonotrophic cycles.

Methods

Mosquito rearing, blood feeding, and processing

\textit{Anopheles gambiae} s.s (Kisumu strain) and \textit{An. arabiensis} (Ifakara strain) mosquitoes were reared under standard insectary conditions of 27 ± 1°C, 70% humidity and a 12-hr light:12-hr dark cycle. \textit{Anopheles gambiae} s.s (Kisumu strain) mosquitoes were provided by Hilary Ranson (Liverpool School of Tropical Medicine). The \textit{An. arabiensis} (Ifakara strain) colony was established in 2008 with individuals from Sagamaganga village (Kilombero District, Morogoro Region, Tanzania)\cite{33}. Larvae were fed \textit{ad libitum} on fish pellets (Tetra Pond Pellets, Tetra GmbH, Herrenteich 78, D49324). Pupae were collected from the larval trays and moved into a cage for emergence. Mosquitoes were considered to be in the age category of “Day 0” on their day of emergence from pupa to adult. Upon emergence, adults were fed \textit{ad libitum} on a 5% glucose solution supplemented with 0.05% (w/v) 4-aminobenzoic acid (PABA).

In order to produce mosquitoes with the same age and different physiological conditions, cages with mosquitoes of the same age (where pupae were added on the same day) were blood fed human blood and membrane feeders at different days after emergence. An oviposition cup was then introduced 2 days after a blood meal to allow egg laying. Mosquitoes were collected 2 days after a blood meal before egg laying (gravid) and 2 days after an oviposition cup was introduced into the cage (sugar fed). Blood feeding was provided to each cage every 6 days. Thus, mosquitoes living 6 or more days after their first blood meal underwent multiple gonotrophic cycles.

Human blood was obtained from the Glasgow and West of Scotland Blood Transfusion Service. Ethical approval for the supply and use of human blood was obtained from Scottish National Blood Transfusion Service committee for governance of blood and tissue samples for non-therapeutic use, and Donor Research (submission Reference No 18–15). Whole blood from donors of any blood group was provided in Citrate-Phosphate-Dextrose-Adenine (CPD-A) anticoagulant/preservative. Fresh blood was obtained on a weekly basis.

Upon collection, mosquitoes were transferred into a cup and killed with a cotton soaked with chloroform placed on top of the cup for 30 minutes. Dead mosquitoes were then transferred into a tube over a layer of cotton and silica gel desiccant. The vial was then immediately stored at 4°C. \textit{Anopheles gambiae} mosquitoes were kept at least one day and \textit{An. arabiensis} at least two days in the vial to allow them to dry completely.

Spectral data acquisition

Individual mosquitoes were laid on their sides on the ATR diamond so that the surface of the diamond was mainly covered by the insect’s head and thorax to avoid as far as possible measuring the contents of the abdomen (Figure 1). The wings and limbs were not removed and were used to help position the mosquito. Pressure was then applied by the anvil of the ATR and the spectrum was measured using a dry-air purged Bruker Vertex 70 spectrometer equipped with a globar lamp, a DLaTGS detector, a KBr beamsplitter, and a diamond ATR accessory (Bruker Platinum ATR Unit A225). 16 scans were taken at room temperature between 400 and 4,000 cm\(^{-1}\) with 1 cm\(^{-1}\) resolution. Mosquito spectra with low intensity, high water content, or a significative atmospheric intrusion (Figure 2) were discarded automatically using \textit{Loco Mosquito} 5.0, a program written in Python 3.6 (see Software availability section).

Machine-learning analysis

A supervised machine-learning approach was used to map the pre-selected 17 wavenumbers to mosquito species (either \textit{An. gambiae} or \textit{An. arabiensis}) and to mosquito age. In both cases, a classification approach was used. Mosquito ages
Figure 1. Best position of the mosquito on the ATR crystal. The correct way to place a mosquito on the ATR crystal (left) is to cover the surface with the head and chest. The wrong way (right) is by centring the abdomen on the crystal.

Figure 2. Common experimental errors during the measurement of the infrared spectrum of a mosquito using ATR-FTIR spectroscopy. Above, blue: Spectrum with a significant atmospheric intrusion. Centre, green: An. gambiae mosquito with high water content. Below, red: Spectrum with poorly defined features due to low intensity, caused by the displacement of the mosquito during the measurement. All spectra are compared to a correct spectrum of a mosquito shown in pink.

1–15 days, taken every two days, which allowed acceptable per- age accuracy while improving on current binary cut-off of 4 days based on oviposition (and assuming no pre-gravid behaviour).

Mosquito species and ages were treated in separate mod-els to increase accuracy. To identify the algorithms most suit-ed to the identification of either mosquito species and age class, we compared the baseline performance of k nearest neighbours (kNN), logistic regression (LR), support vector machines (SVM), random forests (RF), and gradient boosted trees (XGB) using 5-fold cross-validation (Figure 3). This range of parametric (LR, SVM) and non-parametric (kNN, RF, XGB) models offer differ-ent data representation schemes using Euclidian distance (kNN), linear relationships (LR, SVM), and ensemble decision trees (RF, XGB). For species and age class identification, XGB and LR, respectively, were selected for further optimization. The full
dataset—comprising 2,536 mosquito spectral features (details in Table 1) and their corresponding species or age labels—was sampled at random to generate a hold-out validation set stratified according to predicted age classes for each species (see below). The remaining samples were then repeatedly (10 rounds) split in random stratified training and test sets (10 folds). Model optimization involved a further 70%/30% random stratified splitting scheme on each of the training folds.

Each model’s accuracy was calculated against the corresponding test set. The 100 resulting trained models were then ranked according to their accuracy scores, and the 10 best retained and predictions bagged for evaluation of their predicted labels (age or species) against the true labels. All machine learning was performed in Python 3.6 using scikit-learn 0.19, XGBoost 0.82, and corresponding plotting using seaborn 0.9.

Age-structure modelling
To illustrate the utility of our approach for field-based surveys of Anopheles populations, and to assess whether they could be used to measure the impact of vector control interventions in the field, we simulated age structures of An. gambiae and An. arabiensis using a simple age structure population model. Here, age corresponds to days. Specifically, the number of mosquitoes \( N \) surviving to from age \( t \) to \( t+1 \) was modelled as a binomial function: \( N_{t+1} \sim \text{binomial} (N_t, s) \); where \( N_t \) is the total number of mosquitoes alive at age \( t+1 \) and \( s \) is the probability of daily survival. The daily survival rate was based on literature values, i.e., for An. gambiae \( s = 0.91^{56} \) and for An. arabiensis \( s = 0.82^{16} \). For the age structure of the populations under an intervention regime, we assume that the intervention quadruples the mortality rate of both species from day 3 onwards. This emulates a scenario where mosquitoes encounter an insecticide-treated bednet for the first time at day 3, when they start feeding.

Each age class was generated by sampling the full dataset in the proportions calculated from the above simulated age-structured populations. A continuous probability distribution was then fitted to the true and predicted discrete age distributions to better generalize our discrete model predictions to an exponentially decreasing age structure using a half-logistic probability function as

\[
    f(t) = \frac{1}{2} \left[ 1 + \left( \frac{t}{t_0} \right)^{-\frac{1}{2}} \right]
\]

The half-logistic distribution is well-suited for fitting survival data\(^{57,58} \). Age distributions were compared using the Kolmogorov-Smirnov statistic on 2 samples, a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.

Estimation of the light penetration distance in a mosquito
The depth of light penetration for ATR measurements depends on the wavelength \( \lambda \) and angle of incidence of light \( \theta \), and on the refractive indices of the mosquito, \( n_2 \), and the ATR crystal, \( n_1 \):

\[
    d = \frac{\lambda n_1}{2 \pi \sqrt{\sin^2 \theta - (n_2 - n_1)^2}}
\]

Taking into account that, according to the specifications of the ATR accessory\(^{59} \), the incidence angle is \( \theta = 45^\circ \), the Sellmeier equation\(^{60} \) for diamond\(^{61} \) (\( \lambda \) in \( \mu m \)):

\[
    n^2 - 1 = \frac{4.3356 \lambda^2}{\lambda^2 - 0.1060} + \frac{0.306 \lambda^2}{\lambda^2 - 0.1750}
\]

and a Cauchy equation \( n(\lambda) = A + B/\lambda^2 \), with \( A = 1.517 \) and \( B = 8.80 \times 10^{-5} \mu m^2 \) for insect chitin\(^{62} \). The results for the MIR region are shown in Figure 4.
Table 1. Number of mosquitoes of each species and status that have been measured.

<table>
<thead>
<tr>
<th>Anopheles arabiensis</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age/days</strong></td>
<td>1</td>
</tr>
<tr>
<td>Gravid</td>
<td>0</td>
</tr>
<tr>
<td>Sugar-fed</td>
<td>42</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anopheles gambiae</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age/days</strong></td>
<td>1</td>
</tr>
<tr>
<td>Gravid</td>
<td>0</td>
</tr>
<tr>
<td>Sugar-fed</td>
<td>160</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>160</td>
</tr>
</tbody>
</table>

Figure 4. Estimated depth of penetration of the ATR evanescent wave in the mosquito sample.

Results
Mosquitoes preparation
A ‘field-friendly’ protocol to kill and store mosquitoes for infrared (IR) spectroscopy was established as described in Supplementary Note 1. In brief, laboratory-reared female An. gambiae and An. arabiensis mosquitoes of different ages and physiological states were killed by exposure to chloroform for 30 minutes. As chloroform evaporates and does not interact with the mosquito cuticle, the IR spectra were not affected by this chemical (Figure 5). This method, also used before, is more practical in the field than killing mosquitoes with CO₂ or by freezing them at -20°C. Dead mosquitoes were then stored in 20 ml transport tubes with silica gel to dry them out. Removal of water from samples is essential, as it uncovers parts of the IR spectrum that would otherwise be hidden by the intense IR absorption of water (Figure 6). Water IR absorption bands disappeared from An. gambiae and An. arabiensis mosquitoes after storage with silica gel at 4°C for one and two days, respectively (longer in a An. arabiensis due to its higher body water content). In addition, this drying method preserved mosquitoes from decomposition for more than 10 days (Figure 7). Alternative drying methods such as desiccating...
Figure 5. Mid-infrared absorption spectra of a typical mosquito (*An. gambiae*, gravid, 9 days old, top) and liquid chloroform (bottom). Note the absence of the signal of the chloroform employed to kill the mosquito in the insect spectrum, since chloroform rapidly evaporates from the sample and leave no MIR-detectable signals.

Figure 6. Mid-infrared absorption spectra of a recently killed mosquito (blue), a mosquito dried in a vial with silica (green) and in an oven at 80°C for 60 minutes (pink). All mosquitoes were *An. gambiae*, sugar-fed and 11 days old. A clear loss of detail can be observed in the oven dried sample due to heating.
specimens in an oven at 80°C were shown to affect IR spectra, disrupting specially the peaks associated with lipids (Figure 6), and therefore not used.

Spectroscopic method

The far-(30–400 cm⁻¹), mid- (400–4,000 cm⁻¹), and near-infrared (4,000–10,000 cm⁻¹) regions of mosquito spectra were compared (Figure 8). The far- and near-infrared regions were essentially featureless in dried mosquitoes, unlike the NIR spectra previously published³⁸,⁴¹,⁴³,⁴⁵,⁶⁵ which show the intense signals of liquid water when specimens were not dried (Figure 9). However, the mid-infrared region showed a large number of well-defined intense peaks, which are easily identifiable as coming from the chemical components of the cuticle (Table 2). Three different IR spectral sampling techniques were investigated: diffuse reflectance, transmission, and attenuated total internal reflection (ATR, see Spectral data acquisition in Methods). ATR spectroscopy produced the best-defined and most reproducible spectra in the mid-IR region (Figure 10). ATR also allowed the measurement of different parts of the mosquito body (e.g., head or abdomen) that have slightly different IR spectra (Figure 11). It also had superior signal-to-noise ratios allowing acquisition of the spectra in 45 seconds. Raw spectra data is available as Underlying data⁶⁶.

It was estimated that by using the ATR sampling technique in the mid-IR, the light penetrates the sample by 3–10 μm up to about 1000 cm⁻¹, and then up to 22 μm within 1000 and 400 cm⁻¹ (see Estimation of the light penetration distance in a mosquito in Methods). As the cuticle of a mosquito is approximately 2–5 μm thick²¹,⁷, the measured spectra encompass the outer shell and part of the interior of insects. As the cuticle is mainly composed of chitin, proteins, and lipids, spectra associated with these substances were individually compared with the whole-mosquito spectra (Figure 12) to allow the assignment of the main vibrational modes of the mosquito cuticular constituents to each element (Table 2). As the cuticular chemical composition is known to change with species and age⁶⁸,⁶⁹, so too are the relative magnitudes of these vibrational bands. To quantify this change, 17 wavenumbers in the MIR spectrum were selected corresponding to 13 well-defined vibrational absorption peaks (contributed in different proportions by the three main constituents) and 4 troughs (that provide information on spectrum intensity and offset). These 17 wavenumbers were then used for training machine learning models (see below).

Mosquito species determination

To develop a MIRS-based method to determine the age and species of An. gambiae and An. arabiensis, mosquitoes were reared under laboratory conditions (see Mosquito rearing, blood feeding, and processing in Methods) and collected at ages ranging from 1 to 17 days. To model part of the variability typical in the wild, female encompassing a range of physiological states were incorporated in analysis including those that have just taken a blood meal (blood fed), those that had eggs developed in the abdomen (gravid), or that laid eggs but have not blood-fed yet again (sugar fed); mosquitoes undergone either single or multiple gonotrophic cycles depending on their age. In most cases, over 40 mosquitoes per age and physiological condition from each species were analysed (Table 1).
Figure 8. Typical near- (left, blue), mid- (centre, green), and far-infrared (right, pink) spectra of an An. gambiae mosquito. The near-infrared spectrum was collected using diffuse reflectance infrared spectroscopy, while the mid- and far-infrared spectra were obtained using ATR.

Figure 9. Near-infrared diffuse reflectance spectra of water (blue), an undried An. gambiae mosquito (green), a dried mosquito (pink), and chitin (red).
Table 2. Assignment of the wavenumbers assigned in Figure 12\textsuperscript{19,19}.

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>Bond</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3856</td>
<td>*</td>
<td>O-H</td>
</tr>
<tr>
<td>3400</td>
<td></td>
<td>Mosquito moisture</td>
</tr>
<tr>
<td>3276</td>
<td>N-H</td>
<td>Chitin, proteins</td>
</tr>
<tr>
<td>2923</td>
<td>C-H\textsubscript{2}</td>
<td>Proteins, waxes</td>
</tr>
<tr>
<td>2859</td>
<td>C-H\textsubscript{2}</td>
<td>Proteins, waxes</td>
</tr>
<tr>
<td>1901</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>1746</td>
<td>C=O</td>
<td>Proteins, waxes</td>
</tr>
<tr>
<td>1636</td>
<td>C=O</td>
<td>Proteins, chitin</td>
</tr>
<tr>
<td>1539</td>
<td>O=C-N</td>
<td>Proteins, chitin</td>
</tr>
<tr>
<td>1457</td>
<td>C-CH\textsubscript{3}</td>
<td>Wax, proteins</td>
</tr>
<tr>
<td>1307</td>
<td>C-N</td>
<td>Proteins, chitin</td>
</tr>
<tr>
<td>1154</td>
<td>C-O-C</td>
<td>Chitin, waxes</td>
</tr>
<tr>
<td>1076</td>
<td>C-O</td>
<td>Chitin</td>
</tr>
<tr>
<td>1027</td>
<td>C-O</td>
<td>Chitin</td>
</tr>
<tr>
<td>880</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>526</td>
<td>C-C</td>
<td>Proteins, chitin</td>
</tr>
<tr>
<td>401</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

\* Wavenumbers selected as indicators of overall spectra intensity and offset.

Figure 10. Typical ATR (blue, scaled Abs x32), diffuse reflectance (pink), and transmission (green) mid-infrared spectra of a mosquito. The transmission spectrum was taken using ZnSe windows. Its vertical offset is due to the reflection of a part of the light because of the difficulty in controlling the angle of the cell windows with the mosquito inside. The ATR, diffuse reflectance, and transmission spectra are the result of an average of 16, 120, and 80 scans, respectively.
Figure 11. Mid-infrared absorption spectra of the head and thorax (blue) and abdomen (green) of a sugar-fed, 17-day-old, An. gambiae mosquito.

Figure 12. Typical mid-infrared spectrum of an Anopheles mosquito. Shown are An. gambiae (gravid, 9 days-old, blue) and its main chemical constituents wax (arachidyl dodecanoate, green), chitin (from shrimp shells, red), and protein (collagen from bovine Achilles tendon, pink). The wavenumbers selected for the machine learning are indicated with a grey line (Table 2).
A total of 1,522 An. gambiae and 1,014 An. arabiensis spectra from different ages and physiological conditions were used to train supervised machine-learning models (see Machine-learning analysis in Methods). Five algorithms were tested on the dataset to predict mosquito species (Figure 3A). This initial approach identified logistic regression (LR) as the most accurate approach. We generated 100 bootstrapped models which, when aggregated (bagged), predicted the species identity of An. gambiae and An. arabiensis with 76.8 and 76.6% accuracy, respectively. To increase the accuracy of the prediction while retaining the stability and generalisability afforded by bagging, the 10 best models among them were selected, which achieved 82.6% accuracy (Figure 13). These results demonstrate that the MIRS signal is strongly indicative of mosquito species and can be used to distinguish between species in a more time and cost-efficient method, although currently with less accuracy, than standard PCR methods.

Mosquito age determination

After the development of the species-prediction model, a similar supervised machine-learning approach was used to model the chronological age for a given mosquito species. Mosquitoes were screened every second day after emerging as adults, and models trained on the same set of 17 wavenumbers as above. The LR model again performed best for both species in correctly mapping wavenumber intensities to mosquito age (Figure 3B and C). To train, optimise, and validate the models, the full dataset was partitioned into an age-structured validation set and retained for later use in population models (see below). The remaining samples were then randomly split into stratified 70%/30% training and test sets for model tuning. The accuracy in predicting each chronological age varied over mosquito lifespan as in correctly mapping wavenumber intensities to mosquito age (Figure 3B and C). To train, optimise, and validate the models, the full dataset was partitioned into an age-structured validation set and retained for later use in population models (see below). The remaining samples were then randomly split into stratified 70%/30% training and test sets for model tuning. The accuracy in predicting each chronological age varied over mosquito lifespan.

Consistent with natural mosquito populations, but unlike our training dataset, field sampling would not produce age-balanced sample sizes, but rather diminishing sample sizes at older age classes. Furthermore, it would be highly desirable to use our models to measure the impact of vector control interventions on mosquito-population age structures. However, because no real datasets of a true mosquito population age structure exist, the age structures of An. gambiae and An. arabiensis were modelled based on their reported average daily mortality and assuming an intervention that increased the mortality of adult females four-fold after a first blood meal (~3 days after adult emergence).

In these simulations, a starting population of 1,000 female mosquitoes was used, with the population at each subsequent day being calculated as a proportion of the previous day, with survival rates for each species estimated from reports on field.
Figure 14. Prediction of An. gambiae (A–B) and An. arabiensis (C–D) age class using mid-infrared spectra. (A, C) Violin plots of the distribution of per age class prediction accuracies of 100 optimised models. (B, D) Confusion matrices showing the proportion of accurate (diagonal) classification of mosquitoes as either 1, 3, 5, 7, 9, 11, or 15 days old using the 10 best logistic regression models trained on repeated stratified random subsets using 70% of all mosquitoes sampled, and tested on the remaining 30% (n = 681 for An. gambiae and n = 737 for An. arabiensis).
Figure 15. Box-whisker plot containing the measured absorption in each wavenumber and for each age for all the mosquitoes (An. arabiensis and An. gambiae). The orange lines represent the median absorbance of each age and wavenumber, the limits of the boxes correspond to the interquartile range (IQR) and the whiskers show the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.
Figure 16. Reconstruction of the age structure of simulated populations of *An. gambiae* and *An. arabiensis* mosquitoes sampled from simulated pre- and post-treatment populations. Population age structures of *An. gambiae* (A–C) and *An. arabiensis* (D–F) were generated using an age structure population model assuming survival rates of 0.91 (*An. gambiae*, A) or 0.82 (*An. arabiensis*, D), under two common scenarios: naive untreated populations (blue lines), and populations in which a simulated vector control program resulted in 4x daily mortality of mosquitoes after day 3 (see Age-structure modelling in Methods for details). The proportions of each age class were extracted from those simulated populations (A, D), and used to build datasets that are representative of a field-sampled population survey (grey bars in B, C, E, and F). The resulting age-structured dataset was then used as the test set for our age-predicting machine learning models (see Figure 3) and compared with the predicted age structure generated from those models (orange bars in B, C, E, and F). Finally, we fit a continuous probability distribution to the true (grey curve) and predicted (orange curve) for better generalization of our discrete model predictions to an exponentially decreasing age structure. Population distributions were compared using a 2-sample Kolmogorov-Smirnov test (KS_2samp), reported in the y-axis labels. A - Relative proportion of each age class in a simulated population of *An. gambiae*. B - Estimation of age structure of simulated population from (A) using best models from Figure 3B for *An. gambiae* (n = 130). C - Estimation of age structure of simulated population post-intervention from (A) using best models from Figure 3B for *An. gambiae* (n = 122). D - Relative proportion of each age class in a simulated population of *An. arabiensis*. E - Estimation of age structure of simulated population from (A) using best models from Figure 3D for *An. arabiensis* (n = 42). F - Estimation of age structure of simulated population post-intervention from (A) using best models from Figure 3D for *An. arabiensis* (n = 45).
spectrum, this method facilitates prediction of mosquito species distribution and survival, two crucial tasks critical to implement and assess malaria control strategies. An advantage of this approach is that in comparison to the current most widely used technique based on dissection, it can determine the whole-age distribution of a mosquito population from the day of emergence until two weeks of age. Although the accuracy of age prediction in the “mid range” of mosquito life span was not high; this method could accurately estimate the proportion of mosquitoes within the older and most epidemiologically-important age classes that responsible for malaria transmission.

The use of the mid-infrared spectral region provides some advantages over techniques using near-infrared. Foremost, it is possible to independently quantify the amount of different biochemical components as their vibrational bands appear at different wavenumbers. Furthermore, the MIRS bands are more intense and have much greater definition. In contrast, the near-infrared spectrum of a mosquito is composed of few weak signals (Figure 8) that are typically dominated by the much stronger vibrational overtone and combination bands of water (Figure 9)²⁴, which is likely more dependent on the mosquito physiological state and environmental conditions than on other mosquito traits, such as species and age.

We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age with a high degree of accuracy, and ultimately reconstruct population age structures of two important malaria vector species under simulated conditions of changing mortality risk due to vector control. Our algorithms accurately reconstructed age structures of both An. arabiensis and An. gambiae, and also detected shifts in mosquito age structure consistent with simulated impacts of interventions. To our knowledge, this is the first time that a proposed technique has had the ability to predict the age structure of a population or reconstruct demographic patterns as anticipated from specific vector control interventions. The ability of MIRS to predict the changes in mosquito mortality suggests that this approach could constitute an efficient tool for monitoring the efficacy of vector control interventions. Future work will include larger datasets used for training in supervised machine learning, comprising field samples with different ecological conditions. The ecological variability of field samples has limited the use of NIRS for age prediction in wild mosquito populations⁴¹. While the accuracy of MIRS-based approaches may also decline when moving from laboratory-reared to field mosquitoes, we predict that this method will be more robust due to the specific information content and high signal clarity that is obtained in spectra from MIRS. Additional improvements are anticipated by increasing the size and variability of the training set on which mosquito age predictions are validated. This will also facilitate the use of alternative machine learning techniques such as neural networks⁴³ which may yield even higher accuracy and repeatability.

We have shown that MIRS can discriminate between morphologically identical An. gambiae s.l. species with ~83% accuracy. While the observed accuracy of MIRS species prediction is still not comparable to the PCR precision, further work including a larger training set and field samples is expected to increase the overall accuracy of this approach. In addition, the inclusion of other species of the An. gambiae s.l. complex will be necessary to implement this technique for field application. However, these laboratory-based results, which included mosquitoes from different ages, physiological conditions, and cohorts, suggest that despite the ecological and life-history traits variation, MIR spectra contain a species-specific signature that the machine-learning algorithm can detect. Indeed, mass-spectrometry studies have shown that different species in the An. gambiae s.l. complex have quantitative differences in the cuticular hydrocarbon composition of their cuticle⁴⁶, which will affect the MIR spectra.

The biochemical signature obtained by MIRS from the mosquito cuticle provided information on both mosquito species and age. It may therefore be possible to obtain further information on other mosquito traits that alter the cuticular composition. Recently, a new insecticide resistance mechanism has been discovered in An. gambiae, which relies on an increased cuticle thickness that in turn reduces insecticide uptake³⁷. While this mechanism has been detected by electron microscopy, there are no other methods to measure this new trait, which could have profound epidemiological consequences. In the future, MIRS calibrations including cuticular resistant mosquitoes may be able to identify this insecticide resistant trait. In addition, infection with the Plasmodium malaria parasite might be detected by MIRS. Pathogen infection is known to alter mosquito physiology and could directly or indirectly modify their cuticular composition. For example, in the dengue and Zika vector Aedes aegypti mosquitoes, an infrared spectroscopy method has recently been developed to detect Zika virus⁴⁸, the bacterial endosymbiont Wolbachia⁵⁰⁵¹, and malaria infection in mosquitoes⁵₂⁵₃.

The accuracy and generalisability of the MIRS approach presented here shows that this tool holds promise for use in the evaluation of vector control interventions. The inclusion of new species, larger sample sizes and field samples with variable ecological conditions is a prerequisite for the application of this technique. It is worth noting that the cost of a portable FTIR MIR spectrometer is ~$20-25,000, which is in the range of quantitative PCR machines used for species determination and/or insecticide resistance monitoring. However, in contrast to PCR analysis, no additional, ongoing costs for reagents and running costs are required once the core equipment is installed. Thus, this approach could be particularly valuable in resource limited settings.

The MIRS method presented here provides rapid and accurate information on Anopheles species (82.6%) and reliably characterises mosquito age distribution. However, these results were obtained by training machine learning models with a relatively modest number of mosquitoes (2,536). In future work, it will be possible to generate much larger MIRS datasets and thus train more sophisticated predictive models. Such larger data sets will lend themselves to analysis by more powerful “big data” approaches including deep learning methods that would be
expected to improve accuracy considerably beyond this proof-of-principle study. Furthermore, the technique applied to malaria vectors here could also be expanded to other vector-borne diseases such as Zika, dengue, Lyme disease, leishmaniasis, or filariasis. In light of these opportunities, we recommend this method be prioritised for further evaluation.

**Data availability**

**Underlying data**


This project contains the following underlying data:

- DataMosquitoes.zip (zip file containing underlying spectra data)

**Software availability**

Source code: https://github.com/SimonAB/Gonzalez-Jimenez_MIRS/tree/v1.0

Archived source code: http://doi.org/10.5281/zenodo.2609356

**References**


**Grant information**

This work was supported by the Wellcome Trust through a Intermediate Fellowship in Public Health and Tropical Medicine to FO [102350]. This work was also supported by The Engineering and Physical Sciences Research Council (EPSRC) [EP/J009733/1, EP/K034995/1, EP/N508792/1, and EP/N007417/1] and Medical Research Council (MRC) [MR/P025501/1]. FB is supported by an AXA Research Fund fellowship [14-AXAPDC-130] and a European Molecular Biology Organization (EMBO) Long Term fellowship [43-2014]. MV is funded under the MRC/Department for International Development Cordant agreement, which is part of EU EDCTP2 programme [MR/N015320/1].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Acknowledgements**

We would like to thank Dorothy Armstrong and Elizabeth Peat for assistance with mosquito rearing and maintenance. We would also like to thank Hilary Ranson for providing the Kisumu colony.


Open Peer Review

Current Peer Review Status: ✗ 🎉

Version 1

Reviewer Report 19 June 2019

https://doi.org/10.21956/wellcomeopenres.16586.r35616

© 2019 Churcher T. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas S. Churcher
MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

The manuscript presents the first attempt to use mid-infra-red spectroscopy to determine the age and species of Anopheles gambiae s.l mosquitoes. Novel methods for determining the age of the mosquito population are urgently needed and so the new possible technique outlined here is very much welcomed. The paper contains a great deal of information though would benefit from more clarity at times for both the methods and their rationale. Generally, the paper is written with a positive spin and the discussion could be more honest about the frailties of current results (specifically the accuracy) without diminishing the potential of the method. Conversely, the decision to classify mosquitoes into age bands instead of a continuum and the way the results are presented may make the technique look less accurate than it may well be. For example, < 20% of 3 day old A. gambiae mosquitoes are classified as being this age, though the confusion matrix appears to show that it is correctly identifying them as being <7 days (though it is hard to see). This level of accuracy could therefore be tolerated depending on the question under investigation, though people might look at the <20% accuracy figure and write it off.

Specific points are detailed below. I am not qualified to comment on technical aspects of spectroscopy.

Major points:

1. Methods. After being killed “Anopheles gambiae mosquitoes were kept at least one day and An. arabiensis at least two days in the vial to allow them to dry completely.” The two species were therefore, on average, treated differently. Any conclusions about the ability of MIRS to speciate within the gambiae s.l. complex should therefore be tapered as some of the accuracy maybe because the machine learning was picking up differences in time since death/dryness?

2. Methods. Some spectra with low intensity or high water content were removed by the Loco Mosquito 5.0 program automatically. This is available on zenodo yet no criteria for removal were given making it hard to understand what is going on. Did this preferentially exclude mosquitoes of a certain class (you can imagine water content might be related to age)?
3. Methods. Seventeen pre-selected wavelengths were used in the analysis and all other data was excluded. Why this number? It seems strange when there may be considerable information lost. Machine learning methods are designed to deal with this complexity and, though I agree about the dangers of over-fitting, this could be accounted for.

4. Methods. The machine learning methods classified mosquitoes into 7 ages classes. Why was this chosen over a continuous estimation of age? Some justification would be nice as it is different to what is typically done and indeed the authors in the results suggest there was overlap between the bands and training on daily scans (as will happen in the field) made the accuracy worse again. How was this number of classes decided upon? Previous published work has classified mosquitoes as young and old, which would have made the method look much better (it appears from the confusion matrix, it is hard to tell from the shading (consider putting numbers in)). Why do those selected not cover the whole range of ages needed? Figure 16A suggests considerable numbers might be older than 15 days which would be important to the average age estimates.

5. Methods “Mosquito ages 1–15 days, taken every two days”. This does not appear to be the case as no data presented from day 13. Why have different boundaries between the oldest two classes? Is the accuracy of predicting 15 age old mosquitoes only better because there are no other classes within 4 days of them when the others have potentially 4 classes?

6. Results. The language of the results sometimes doesn’t come across as very balanced and veers off into opinion. For example, the statement “These results demonstrate that the MIRS signal is strongly indicative of mosquito species and can be used to distinguish between species in a more time and cost efficient method,” not everyone would agree with given the accuracy.

Minor points:
1. Abstract. Species accuracy is given as 82% when it only achieved an average of 76% using the steps outlined in the methods. Dropping 90 of the worst performing models gets you to the better value, but this is only done post hoc. You should either say you are going to do this in the methods or change the headline value in the abstract as it is currently confusing when the reader looks at the abstract and then Figure 13A.

2. Penultimate paragraph page 4. “16 scans were taken at room temperature”. Why was this number done and were they averaged for the spectra used in the analyses?

3. There are lots of methods in the results section. Some of this is repeated and some isn’t, the former could be deleted but the later could be transferred to the methods to aid the reader.

4. The discussion states: “this method could accurately estimate the proportion of mosquitoes within the older and most epidemiologically-important age classes that responsible for malaria transmission”, though mosquitoes <15 days of age are likely to be infectious. Suggest revising.

5. The discussion rather grandly says “To our knowledge, this is the first time that a proposed technique has had the ability to predict the age structure of a population or reconstruct demographic patterns as anticipated from specific vector control interventions”. I’m not sure this is true, as other techniques have showed the ability but the real proof of the pudding comes when moving to the field with the considerable difference between mosquitoes (which the author highlights). Consider revising.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
In the hope to improve this article, here are my suggested edits:

In the abstract, title, or keywords there is no mention where this study is conducted. This is significant information because depending on the part of Africa, different members of *Anopheles gambiae* complex exists so the utility of testing done here distinguishing *Anopheles gambiae* from *An. arabiensis* may not transfer well in other regions where *An. gambiae s.s.* and *An. coluzzii* coexists. It is up for further study but the manuscript should be clear on the limited scope of testing done for this study to help readers contemplating using this method.

Abstract “The accuracy of classifying mosquitoes by species was 82.6%” -> Please provide variation or standard deviation of this estimate. Also the Figure 3 data shows all less than 80% so I don’t know how the average could be 82.6%. What about the number for age prediction accuracy? Also state which model was best performing in predictions and the provide correct assignment probability for that in the abstract as author explored multiple models.

“a negligible cost” -> provide an actual estimation in numbers (dollar or euro amount) and what’s included in the cost (equipment, labor, supplies).

“a promising alternative to current mosquito species” -> Many would disagree that 82.6% accuracy in species diagnostic is a promising alternative. I can see the method could be a triage method for selecting specie of interest when dealing with large number of samples for genetic screening.

First paragraph of introduction – because the method introduced here is nothing to do with insecticide resistance, this paragraph containing lengthy intro about the involvement of insecticide resistance seem unnecessary and irrelevant. The second is direct to the point of this paper and good and succinct starting point.

“Currently, *An. gambiae* s.l. species are best distinguished by polymerase chain reaction (PCR) methods” -> The citation 33 is no longer valid because it doesn’t distinguish *An. gambiae s.s.* from *An. coluzzii*. The methods by Favia et al. (2001), Fanello et al. (2002), and Santolamazza et al. (2008) would be more appropriate for citation here.

“polymerase chain reaction (PCR) methods, which are time-consuming and expensive”. It is relative to call this method “expensive” as it would cost $2-3 US dollars to screen a single sample as opposed to Lee et al. (2015) which cost $5-6 USD for a single sample but collects a wide variety of epidemiologically important traits. Of course, SNP chip or sequencing approach would cost in order of several hundreds per sample. So this molecular assays cited are on the cheap side already. Perhaps more accurate to say “which are time-consuming and still relatively costly, and can thus only be carried out on a subsample of mosquitoes collected during typical entomological surveillance conducted by many agencies in Africa”

Method section mosquito rearing – it is not clear where the mosquito rearing occurred. Is it in Tanzania or Glasgow? Specify the location of *An. arabiensis* Ifakara strain was established.

Method section “Mosquitoes were collected 2 days after a blood meal before egg laying (gravid) and 2 days after an oviposition cup was introduced into the cage (sugar fed).” – This part is confusing.

“mid-infrared spectroscopy-based prediction of mosquito age structure was ….” -> Capitalize the first letter of the sentence.
At the end of 2nd paragraph of introduction: “…the assessment of the impact of these and other control measures”. -> It is not clear what “these and other” control measures refer to. I think it is better to drop the first paragraph and then change “these and other” to “various vector”

“The An. gambiae s.l. complex includes several cryptic species that can only be distinguished by molecular analysis” - cryptic species implies two or more species hidden under one species name. I would cite the typical species diagnostic tools only dealing with known “good” species, I would use ‘sibling’ instead of ‘cryptic’.

The introduction paragraph starting with “As in the case of age determination…” seems out of place. 3rd and 4th paragraphs talk about age determination, 5th paragraph talks about species determination and 6th paragraph going back to age determination. Some rearrangement of the text necessary to make the story flow better. This paragraph is also very lengthy and makes readers wonder if there is an aging method already available for Anopheles species using NIRS. Rather than lengthy and detailed description of what NIRS and why it has problems, it seems the paper is better served by introducing the method of choice and its strength with appropriate citations. To be specific, I would remove the text starting with “while promising, …” till the end of 6th paragraph.

“Here we tested if these limitations can be overcome by shifting the measurement range (25,000-4,000 cm\(^{-1}\)) to the mid-infrared region (4,000-400 cm\(^{-1}\)), employing an attenuated total reflectance (ATR) device to assess the mosquitoes, and modelling the results with supervised machine learning.” -> In relation to the edits suggested to the previous paragraph, I would change this to “Here, we tested the id-infrared region (4,000-400 cm\(^{-1}\)), employing an attenuated total reflectance (ATR) device to assess the mosquitoes which is relatively small.” This suggested edit would make the following phrase in the sentence “Modelling the result with supervised machine learning” awkward. However, the next paragraph introduces why this approach was used so this particular phrase can be removed from this paragraph. Also any reason that the numbers are from large to small?

Method Spectral data acquisition section - “Individual mosquitoes were laid…” -> change it to “dried specimens” to make the experimental process clearer to the readers.

“a globar lamp” -> G should be capital on globar.

“DLaTGS “ -> spell it out like “Deuterated Lanthanum α Alanine doped TriGlycine Sulphate (DLaTGS)”

“KBr beamsplitter” -> spell it out like “Potassium bromide (KBr)”

“Bruker Platinum ATR Unit A225” -> add company name and location of the company

Method Spectral data acquisition “4,000 cm\(^{-1}\)” -> “-1” should be superscript.

“Loco Mosquito 5.0” – add a text like “a custom program” to indicate that this program is developed by the authors.

Figure 1. I would prefer if the figure could be self-explanatory as much as possible without having to read legend. Add text or symbol in the figure above each sub-figures what is “correct” and “wrong” way.

Figure 2. It is best if top panel includes an example of “good” reading to differentiate the bad readings. Also I would add text on each panel like “atmospheric intrusion”, “high water content”, and “low intensity”. Pink and red are hard to distinguish. Perhaps dotted lines for pink?
“mod-els” -> “models”

“suit-ed” -> “suited”

“ma-chine” -> “machine”

References

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Medical entomology, population genetics, malaria vectors, bioinformatics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.