Time-Series Mining Approaches for Malaria Vector Prediction On Mid-Infrared Spectroscopy Data

RESEARCH PAPER

LUCAS G. M. CASTRO

HENRIQUE V. COSTA

VINICIUS M. A. SOUZA

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

Malaria is an infectious disease caused by the *Plasmodium* parasite transmitted to humans by the bite of infected female *Anopheles* mosquitoes. The disease remains a major cause of child mortality globally and caused more than 600,000 deaths in 85 countries only in 2022, predominantly affecting the African region. Recent discussions point out that climate change is expanding the geographical distribution of mosquitoes, accelerating the malaria burst in areas free from outbreaks. Traditional vector control relies on chemical methods (e.g., insecticides), but effective control implementation requires accurate and cheap mosquito population monitoring and longevity estimates. This study investigates using mid-infrared spectroscopy (MIRS) data as input for efficient time-series classification methods to predict the species and age of malarial mosquitoes. Unlike previous studies using traditional machine learning, our comprehensive evaluation includes 14 algorithms from four time-series mining approaches, such as feature-based, interval-based, convolutional-based, and deep learning methods. These methods consider the particularities of time series, such as temporal dependencies and correlations between observations. Results indicate that the deep learning algorithm InceptionTime achieves 97% species identification accuracy and 83% age prediction accuracy, outperforming the traditional methods. This research contributes to the field by highlighting the effectiveness of time-series mining approaches for malaria vector control using spectroscopy. As malaria continues to pose a significant threat, these advancements contribute to developing innovative and efficient tools for malaria control strategies.

CORRESPONDING AUTHOR:

Vinicius M. A. Souza

Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná Rua Imaculada Conceição 1155, Curitiba, PR, Brazil

vinicius.mourao@pucpr.br

KEYWORDS:

mosquitoes; malaria; time-series classification; mid-infrared spectroscopy; machine learning; deep learning

TO CITE THIS ARTICLE:

Castro, LGM, Costa, HV and Souza, VMA 2024 Time-Series Mining Approaches for Malaria Vector Prediction On Mid-Infrared Spectroscopy Data. *Data Science Journal*, 23: 25, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2024-025

1 INTRODUCTION

Malaria is one of the leading causes of death in children. According to data from the Institute of Health Metrics and Evaluation (IHME), malaria is responsible for 12% of all child deaths.[1] Globally, in 2022, there were an estimated 249 million malaria cases and 608,000 deaths in 85 endemic countries (WHO 2023). Malaria disproportionately affects the most marginalized populations. In 2022, the Africa region was home to about 94% of all malaria cases (233 million) and 95% of deaths (580,000). Four countries accounted for just over half of all malaria deaths worldwide: Nigeria (26.8%), the Democratic Republic of the Congo (12.3%), Uganda (5.1%), and Mozambique (4.2%) (WHO 2023).

This infectious disease spreads to humans through the bite of a female infected mosquito of the *Anopheles* genus. The mosquito transmits a parasite of the *Plasmodium* group that enters the victim's blood system and travels into the person's liver, where the parasite reproduces. Once in the liver, the *Plasmodium* parasite undergoes further development, transforming into a form capable of infecting red blood cells. This intricate life cycle within the human host leads to a recurrent cycle of symptoms, including fever, chills, and anemia. The infection can become severe, causing organ failure and, in extreme cases, leading to death if not promptly diagnosed and treated.

Mosquitoes' reproduction and spatial distribution of adults are strictly related to climatic factors that can influence the behavior and survival of species (Cella et al. 2019). Mosquitoes are prevalent in tropical and subtropical regions. However, the increases in temperature, humidity, and rainfall caused by climate change are helping to proliferate the mosquito population at higher altitudes regions (Caminade et al. 2014). This environmental change is widening the geographical distribution of malaria, allowing it to emerge in new locations that had previously not supported mosquito populations (Lubinda et al. 2021). Additionally, increases in temperatures at lower altitudes, where mosquitoes and malaria are already prevalent, allow it to develop malaria faster and increase transmission rates (Cella et al. 2019; Lubinda et al. 2021).

Combating and controlling malaria requires data-driven strategies and innovative tools for more efficient and effective solutions. As data-driven solutions, we can mention an outbreak early warning system that predicts malaria outbreaks based on climatic factors (e.g., temperature, precipitation, wind speed, solar radiation) in a region using machine learning algorithms (Modu et al. 2017). An example of an innovative tool is the intelligent trap that can capture target species (e.g., *Aedes* or *Anopheles* mosquitoes) according to their wing beat frequency measured by optical sensors (Souza 2017; Souza et al. 2020).

Traditionally, most insect vector control comprises chemical methods (e.g., adulticides applied as space sprays outdoors and indoors or insecticidal nets fitted in houses as curtains or screens) that target adult mosquitoes. These methods intend to reduce the densities, longevity, and biting behavior of mosquitoes (Ritchie et al. 2021). However, the effectiveness of such methods requires understanding the local population's abundance and behavior of the vector mosquitoes. Thus, the accurate estimation of mosquito populations and their longevity is essential for assertive employment of the control measures and evaluating their effectiveness over time. The average age of a mosquito population is the most important determinant of vectorial capacity and the likelihood of disease transmission (Johnson et al. 2020).

Population density estimation depends on identifying mosquito species and age, which requires costly or time-consuming methods. Most methods are based on regularly collecting and manually counting mosquitoes from traps such as CDC light or CO_2 traps (Sriwichai et al. 2015). For age estimation, an entomology expert dissects the mosquito ovaries and examines the presence and number of eggs (Johnson et al. 2020). However, some species, such as *Anopheles gambiae* and *Anopheles arabienses*, two of the most prevalent malaria vectors in Africa, can be distinguished only by molecular analysis due to their morphological similarities (Santolamazza et al. 2008).

Molecular analysis methods, such as Polymerase Chain Reaction (PCR), are time-consuming and can only be carried out on a subsample of insects trapped in a region. Recently, González

[1] https://ourworldindata.org/malaria.

Jiménez et al. (2019) have shown that mid-infrared spectroscopy can be faster and more cost-effective than traditional or PCR-based methods. In this technique, mosquitoes are subjected to the emission of infrared light, and the absorption of this light by mosquito tissue generates spectral data in which machine learning algorithms can recognize complex relationships that distinguish species and age. Each data point in the spectrum corresponds to the intensity of the measured signal at a specific wavelength.

The spectral data generated by the mid-infrared spectroscopy can be considered a time series, that is, a real-valued sequence of continuous data in which the observations are highly correlated. While spectral data from mid-infrared spectroscopy is not a traditional time series in the temporal sense, it is a real-valued sequence of measurements across different wavelengths. The observations within the spectrum can be highly correlated due to the underlying physics of molecular vibrations.

Previous studies using mid-infrared spectroscopy data for species and age prediction of malarial mosquitoes consider traditional machine learning algorithms (González Jiménez et al. 2019; Mwanga et al. 2019; Siria et al. 2022). An essential assumption of these algorithms is that the features of an example are independent. However, the features of time-series data are a sequence of observations that have temporal dependencies and correlations, which means that the value of a variable at one time point is influenced by its past values. Besides, this data have characteristics such as peaks, valleys, trends, and periodic patterns. Thus, a supervised model induced for this data must consider such data particularities. In this direction, we present and investigate the performance of time-series mining algorithms, a category of algorithms that take into account the sequential nature of the observations and the temporal dependencies.

This work comprehensively evaluates the state-of-the-art classification methods for time-series data in malarial vector species identification and age prediction. Specifically, we consider 14 algorithms from four time-series mining approaches: *i)* feature-based, *ii)* interval-based, *iii)* convolutional-based, and *iv)* deep learning. Our results show that the deep learning algorithm InceptionTime is an accurate method able to identify 97% of species correctly and predict the age of insects with 83% accuracy, outperforming the current results from the literature using traditional machine learning algorithms such as a SVM trained with specific wavenumbers from the spectroscopy or convolutional neural networks. Compared with complex deep learning methods, the convolution-based algorithm Rocket presents competitive results with a low computational cost. In addition, our results reinforce the power of mid-infrared spectroscopy data as input for supervised learning algorithms to distinguish species with morphological similarities correctly.

We organized this work as follows: Section 2 discusses the related work. Section 3 introduces the material and methods employed, such as mid-infrared spectroscopy, data description, preprocessing steps, and an overview of state-of-the-art time-series classification approaches and algorithms. The results and analysis are presented in Section 4. The limitations and advantages of each approach are discussed in Section 5. Finally, Section 6 presents our conclusions.

2. RELATED WORK

Researchers have been investigating how to analyze insect behavior and identify species automatically using different techniques, technologies, and approaches for decades. According to Mankin et al. (2011), the first works aimed to identify the presence or absence of insects based on acoustic or vibrational monitoring for pest control, such as proposed by Main in 1909. In 1945, Kahn, Celestin, and Offenhauser used microphones to record the behavior of different insect species and distinguished males from females according to their noise. In 1991, Moore proposed an optical sensor to record the variation of the light caused by the insect crossings through the sensor, being able to automatically identify two species of *Aedes* genus using an artificial neural network. More recently, many studies have focused on improving electronic devices based on LED, laser or infrared light to measure the wing beat frequency of insects for their classification into species using supervised machine learning algorithms (Batista et al. 2011; Fernandes, Cordeiro & Recamonde-Mendoza 2021; Potamitis & Rigakis 2016; Souza et al. 2020).

Although using information from insect wing beats extracted by optical or acoustical sensors is a promising approach for species identification, it is notable that insects' age and environmental conditions, such as temperature and humidity, affect each species differently, making the classification difficult in the field and requiring costly and large datasets built under varying conditions (Parmezan et al. 2021; Souza et al. 2020). Furthermore, these techniques may be efficient for distinguishing species from different genera (e.g., *Aedes* and *Anopheles*), but may not be accurate for distinguishing similar species of the same genus. In this direction, species identification based on molecular analysis is a precise and invariant technique to those conditions. However, molecular analysis techniques such as Polymerase Chain Reaction (PCR) are time-consuming and require repeated supply of reagents, making them expensive and unreliable in poorly resourced settings (Mwanga et al. 2019). Recent studies have shown that non-molecular techniques such as infrared spectroscopy are effective and a cheaper alternative, achieving competitive results to those obtained by molecular analysis. Thus, we discuss the recent advances focusing on MIRS methods for mosquito analysis.

González Jiménez et al. (2019) use mid-infrared spectra of mosquitoes to characterize both age and species of African malaria vector species *Anopheles gambiae* and *Anopheles arabiensis*. The study evaluated four traditional supervised machine learning algorithms: k-Nearest Neighbors, Logistic Regression, Support Vector Machines, Random Forests, and gradient-boosted trees with XGBoost. The algorithms were trained with 17 features regarding specific wavenumbers from the spectra. In the experiments on a dataset comprising 2,536 examples, the authors reported 82.6% accuracy for species identification using Logistic Regression. The authors also predicted the age of each species from 1 to 15 days and reported results ranging from an average of 15% to 97% for *Anopheles gambiae* and 10% to 100% for *Anopheles arabiensis*.

Mwanga et al. (2019) evaluated seven traditional machine learning algorithms (k-Nearest Neighbors, Logistic Regression, Support Vector Machines, Random Forests, XGBoost, Naive Bayes, and MultiLayer Perceptron) to identify blood meal sources in malaria vector using mid-infrared spectroscopy. The algorithms accurately distinguish between vertebrate blood meals in the guts of *Anopheles arabiensis* mosquitoes. The spectra data were used to classify blood meals into one of four host species classes: bovine, chicken, goat, and human. The Logistic Regression classifier obtained the best results with 98.4% overall accuracy, being 96% for goat blood, 97% for bovine blood, and 100% for chicken and human blood. The results were obtained from a dataset with 2,000 examples in which 90% was used for training and 10% for testing.

Siria et al. (2022) employ a deep learning model to predict the age and species of *Anopheles gambiae, Anopheles arabiensis* and *Anopheles coluzzii* mosquitoes using MIRS data. Specifically, the authors consider a Convolutional Neural Network (CNN) composed of five 1D convolutional layers comprising 16 filters each to capture complex local features in the spectra, followed by fully connected layers to capture the correlation of the extracted features across the entire spectra. The model obtains 95.33% accuracy for species identification and 95% for age prediction, not comparing the results against other deep learning algorithms or traditional supervised machine learning. The results were measured in a dataset with spectra from 41,151 female mosquitoes reared in different laboratories in Tanzania and Burkina Faso to guarantee variability.

While mid-infrared spectroscopy presents essential advantages over molecular analysis techniques, the identification of species, sex, and age of insects requires machine learning algorithms tailored for MIRS data. As noted in the discussion of related work, much of the work considers traditional machine learning algorithms that do not address the unique aspects of time series. Thus, the literature lacks a comprehensive comparison that considers suitable methods, as carried out in this work.

3. MATERIAL AND METHODS

3.1 MID-INFRARED SPECTROSCOPY

Spectroscopy is a technique that investigates the interaction between matter and electromagnetic radiation (e.g., light), focusing on the relationship between the radiation's wavelength or frequency and its interaction with the material. The method involves measuring and analyzing the spectrum of absorbed radiation by a substance under analysis.

Infrared spectroscopy is used in entomological surveillance to identify mosquito species by quantifying how their cuticles absorb infrared light. Most works generally consider near-infrared spectroscopy (NIRS) for examining insects, in which the spectrum is restricted from 10,000 to 4,000 cm^{-1} (Barbosa et al. 2018; Johnson 2020; Mayagaya et al. 2009). Although efficient in identifying insect species, age prediction using NIRS is still challenging (Siria et al. 2022). In order to obtain better age predictions, researchers have investigated the use of mid-infrared spectroscopy (MIRS) (González Jiménez et al. 2019). MIRS considers the absorption spectrum in the mid-infrared region from 400 to 4,000 cm^{-1}. In Figure 1, we illustrate the mid-infrared absorption spectra obtained by *Anopheles gambiae* and *Anopheles arabienses* mosquitoes, studied in this work.

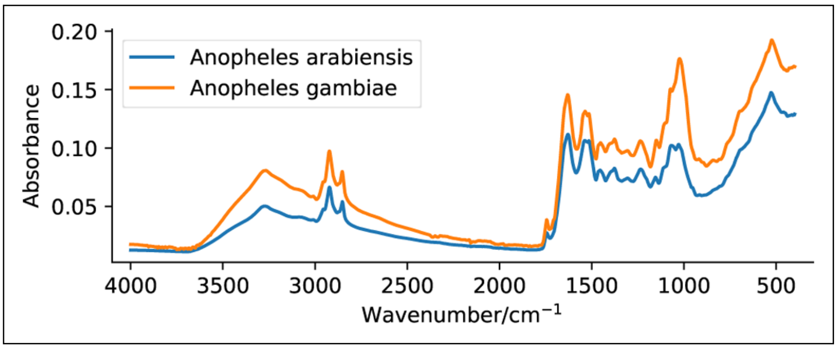


Figure 1 Example of mid-infrared spectroscopy obtained from two insect species.

As we note in Figure 1, the data generated by MIRS are time series, that is, real-valued continuous data with a length of 3,600 observations. These series are used as input for machine learning methods to predict species and age of insects.

3.2 DATA AND PREPROCESSING

Our study considers an MIRS dataset collected by Siria et al. (2022) and publicly available in the Enlighten database.[2] This dataset contains spectra of 41,151 female mosquitoes aged 1 to 17 days from the following malaria vector species: *Anopheles gambiae* (AG), *Anopheles arabiensis* (AA), and *Anopheles coluzzii* (AC). In order to increase variability, the data was collected from mosquitoes with different physiological states and the spectroscopy was measured in laboratories in the UK, Tanzania, and Burkina Faso. From this data, we perform a set of preprocessing procedures to train supervised algorithms from different approaches to predict the species and age of malarial mosquitoes, as discussed in the following.

Initially, we discarded the examples from the *Anopheles coluzzii* species to reduce the class imbalance since this species represents only 3% of the examples, making the steps of training and testing classifiers more difficult. Thus, our experiments consider 39,716 examples from two species, which we split into 70% for training and 30% for testing. Since we have two tasks (i.e., species and age prediction), we split the data into stratified training and test sets according to the labels of each task. For age prediction, we grouped the examples spanning 17 days into three age classes: 1–4, 5–10, and 11–17 days. Figure 2 illustrates the class distribution for both tasks considered in our holdout evaluation.

We have noted that the original data have varying sampling rates, with examples from 1,701 to 14,932 observations due to the different equipment used in the analysis. However, we need a fixed length to use the data as input to train the machine learning models. In this direction, we downsampled all examples to have 1,000 observations after applying an anti-aliasing filter (Broersen & de Waele 2000). Such a dimensionality reduction also contributes to the computational cost spent on model training or feature extraction.

Finally, the last preprocessing step was the normalization to guarantee that all examples are in the same range of values. Without normalization, observations with larger magnitudes or variances may dominate the learning process, leading to biased model training and inaccurate classification results. Normalization also reduces the impact of outliers.

[2] https://github.com/SimonAB/DL-MIRS_Siria_et_al.

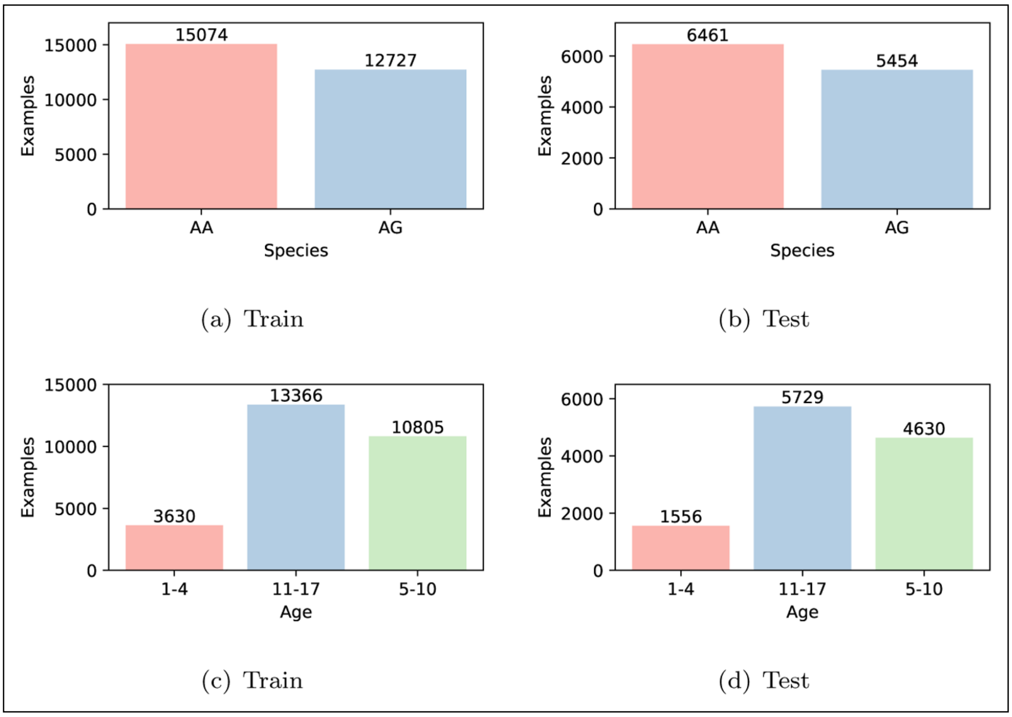


Figure 2 Class distribution into training and test sets for the tasks of species prediction (*top*) and age prediction (*bottom*). For species prediction, AA represents the species *Anopheles arabiensis* and AG represents *Anopheles gambiae*.

In this phase, we employed z-normalization (Lima & Souza 2023). This method rescales the time-series values to have zero mean and standard deviation close to one. The normalized values *X'* of a time series *X* are obtained according to Equation 1, in which μ and σ are, respectively, the mean and standard deviation of a series with *n* observations.

$$X_i' = \frac{X_i - \mu}{\sigma} \tag{1}$$

In Figure 3, we illustrate the same time series previously shown in Figure 1 after resampling to 1,000 observations and z-normalization, responsible for changing both horizontal and vertical original axes.

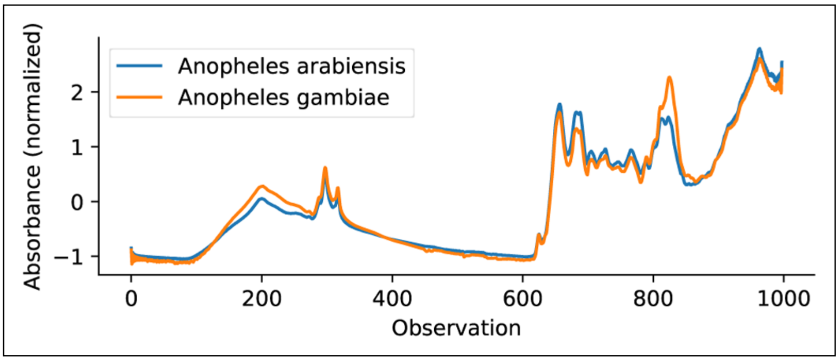


Figure 3 MIRS data after the preprocessing steps of dimensionality reduction and normalization.

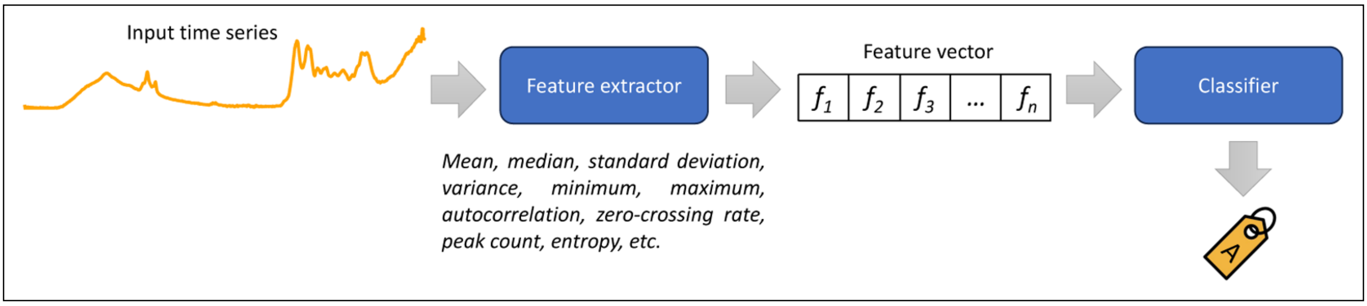
3.3 TIME-SERIES CLASSIFICATION

A time series $X = (x_1, x_2, \ldots, x_n)$, such as the MIRS data previously illustrated in Figure 3, is an ordered sequence of *n* real-values x_i measured at equal time intervals, in which x_i represents a value observed at time *i*. In the classification task, we aim to assign a class label ŷ (e.g., AG or AA) to an unknown query time series using a supervised model built with a labeled training set $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ of *N* time series where Y_i denotes the target variable for each X_i.

Since the features of time-series examples are ordered and correlated, classification models for this data differ from traditional models in which the features are independent. In the following, we introduce the main approaches for time-series classification based on the categories discussed in Bagnall et al. (2017) and other recent proposals (Dempster, Schmidt & Webb 2023; Middlehurst et al. 2021).

3.3.1 Feature-based

Feature-based classifiers consider a preliminary step of extracting descriptive statistics from the time series to use them as features to train conventional machine learning algorithms, such as Random Forests, Support Vector Machines (SVM), and XGBoost. The process is illustrated in Figure 4 given an input time series. The most straightforward approach is to use the raw data comprising the 1,000 observations of the time series as descriptive features for each example.



Mean, median, standard deviation, variance, minimum, maximum, autocorrelation, zero-crossing rate, peak count, entropy, etc.

Figure 4 General process of the feature-based approach.

Another strategy is to consider a subset of values as features instead of all observations. For example, González Jiménez et al. (2019) consider the values observed at 17 specific positions of the MIRS time series that represent the absorption of pre-selected wavenumbers. These wavenumbers contain well-defined intense peaks, which are easily identifiable as coming from the chemical components of the cuticle, being discriminant to identify the species. Figure 5 illustrates the positions of such wavenumbers in an MIRS time series.

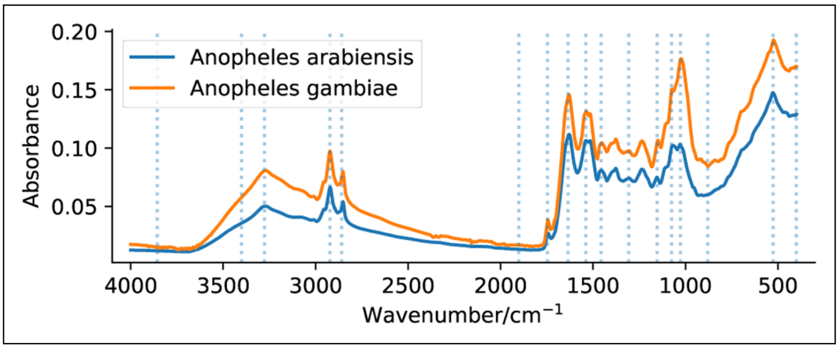


Figure 5 Wavenumbers selected as features to train supervised machine learning algorithms.

For more general time-series classification problems, we can consider statistical features in the time domain such as mean, maximum, minimum, standard deviation, mean absolute deviation, skewness, and kurtosis or structural features such as trend, seasonality, periodicity, and self-similarity. Besides, we can transform the series from time to frequency domain by Discrete Fourier transform (DFT) to extract features such as spectral irregularity, flux, roll-off, and energy (Silva et al. 2015). We can also transform the time series to a two-dimensional representation, such as a Recurrence Plot, then extract statistical descriptive features related to texture (Souza, Silva & Batista 2014).

Highly Comparative Time-Series Analysis (HCTSA) is a library that computes over 7,700 features from time series (Fulcher & Jones 2017). Recently, Lubba et al. (2019) introduced the 22 CAnonical Time-series CHaracteristics (Catch-22), which identified a reduced subset with the 22 most discriminatory features from the HCTSA that provides competitive results for different time-series problems. Among the selected features, we can highlight: *longest period of consecutive values above the mean, time intervals between successive extreme events above and below the mean, first minimum of autocorrelation function, centroid of the Fourier power spectrum, change in correlation length after iterative differencing, periodicity*, among others. It is noteworthy to mention that such features are suitable for the particularities of time series and implicitly consider the temporal dependence of the observations.

In this work, we also combine the 22 features computed by Catch-22 with the 17 values observed at specific wavenumbers, composing a new feature set with 39 descriptive features to train the classifiers.

3.3.2 Interval-based

While feature-based methods extract features from the whole time series, interval-based methods select one or more intervals (i.e., contiguous subsequences) of the series to derive features. An advantage of this approach is to discard regions of the series with noise that could confound the classifier through irrelevant features. For example, in Figure 6, we show five examples from the same class (AG) and highlight two different regions. The region on the left represents a promising candidate interval to extract discriminatory features. On the other hand, the region on the right side contains a significant variation between the examples of the same class, and the extraction of features in this interval can confound the classifier.

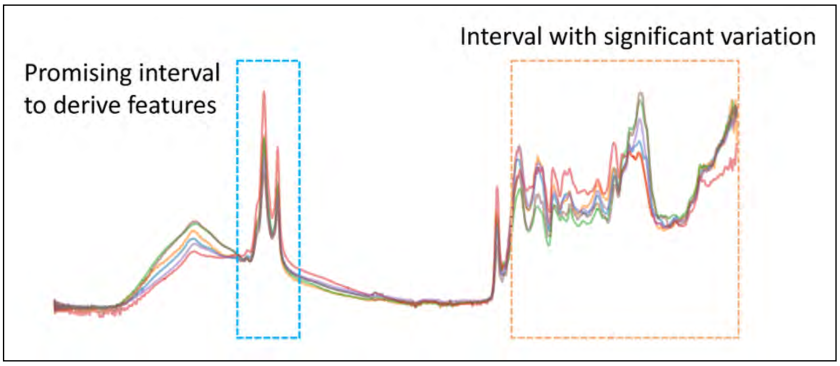


Figure 6 Example of promising and unpromising intervals on MIRS data of *Anopheles gambiae*.

For a given time series with *n* observations, there are $\frac{n(n-1)}{2}$ possible intervals, and a challenge is finding the best interval. In general, these algorithms generate different random intervals and classifiers on each one, ensembling the resulting predictions.

The most representative interval-based algorithms are Time Series Forest (TSF) (Deng et al. 2013), Canonical Interval Forest (CIF) (Middlehurst, Large & Bagnall 2020), and interval-based Diverse Representation Canonical Interval Forest (DrCIF) (Middlehurst et al. 2021).

TSF is an ensemble of tree classifiers based on Random Forest (Breiman 2001), built on the summary statistics of randomly selected intervals from the time series. For each tree, *k* intervals with a random start position and length are randomly selected. The mean, standard deviation, and slope are extracted from each interval and concatenated into a feature vector with 3*k* length. These features are then used to build a tree, which is added to the ensemble. A TSF predicts a testing instance as the majority class according to the votes from all time-series trees in the ensemble. TSF considers 500 trees and $\sqrt{n}$ random intervals by default.

CIF classifier is an adaptation of TSF that embeds the Catch-22 features. The classifier considers the three TSF features along with the 22 features from Catch-22. In order to add diversity to the ensemble, eight of the 25 features are randomly chosen for each tree.

DrCIF is an extension of CIF that extracts features from multiple intervals taken from the original series, the first-order difference series, and the periodograms of the whole series. Seven basic summary statistics are extracted from the interval of any of the three representations: mean, standard deviation, slope, median, interquartile range, minimum, and maximum. DrCIF adds the Catch-22 features to form a set of 29 features. As performed by CIF, eight of the 29 features are randomly selected for each tree.

3.3.3 Convolution-based

The convolution-based methods are computationally efficient classifiers that use many random convolutional kernels to extract features from the time series and use them as input for linear classifiers such as logistic regression. For time series, a kernel is a vector of weights (e.g., [–1, 0, 1]) convolved with a time series through a sliding dot product operation to produce a feature map. Such convolution kernels at different and random lengths and weights can capture complex patterns and shapes to discriminate the series. The main methods of this category are Rocket (Dempster, Petitjean & Webb 2020) and MiniRocket (Dempster, Schmidt & Webb 2021).

Instead of learning a convolutional kernel as performed by convolutional neural networks, Rocket (RandOm Convolutional KErnel Transform) generates more than 10,000 random convolutional kernels that capture relevant features from time series when combined. Each one-dimensional kernel with random weights, length, bias, dilatation, and padding slides through a time series, performing the dot product and producing a transformed time series named feature map. Rocket computes two aggregate features from this transformed time series by a pooling operation: the maximum value (i.e., global max pooling) and the proportion of positive values (PPV). Figure 7 illustrates the whole feature extraction process by convolution considering a single kernel. Since Rocket produces two features per kernel (MAX and PPV) and there are 10,000 random kernels, the algorithm produces 20,000 features per time series. These features are then used to train a linear classifier, such as ridge regression or logistic regression using stochastic gradient descent.

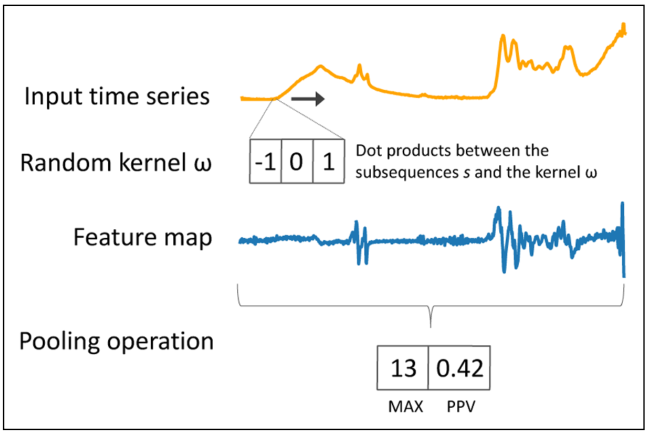


Figure 7 Process of kernel convolution for feature extraction in which two features (MAX and PPV) are extracted from the transformed time series (or feature map). Rocket performs such a process for 10,000 random kernels generating 20,000 features for training a linear classifier.

A limitation of Rocket is the large number of kernels required to make it accurate. Besides, Rocket is non-deterministic due to the use of random kernels. In this direction, MiniRocket (MINImally RandOm Convolutional KErnel Transform) follows the main steps of Rocket, reducing the computational cost and removing the randomness by using a small and fixed set of kernels. MiniRocket uses kernels of length 9, with weights restricted to two values, and computes only PPV from the transformed time series. Thus, the algorithm uses 10,000 features to train a linear classifier.

3.3.4 Deep learning-based

Conventional machine learning algorithms such as SVM or Random Forests learn a predictive model from handcrafted features extracted from data, as discussed in Section 3.3.1, requiring domain experts to design adequate extractors according to the problem. On the other hand, deep learning algorithms consider an end-to-end approach in which the entire predictive task is addressed as a single and integrated system without the need for explicit feature engineering. These algorithms receive raw data as input and generate the answer as output after transforming the data into multiple levels of representations that contain relevant and discriminative features (LeCun, Bengio & Hinton 2015). In general terms, a deep neural network is a composition of multiple hierarchical parametric functions (in practice, the layers of a network) where each layer is a different representation of input data (Sarker 2021).

While deep neural networks, commonly referred to as deep learning, have a well-established track record in computer vision applications like face recognition (Lawrence et al. 1997) and object detection (Zhao et al. 2019), their application to one-dimensional data as time-series classification problems has only gained prominence in recent times, as demonstrated by the competitive results discussed in Ismail Fawaz et al. (2019) in a comprehensive comparison considering time-series benchmark datasets from different domains.

Currently, InceptionTime Ismail Fawaz et al. (2020) is the state-of-the-art neural network for time-series classification. InceptionTime is an ensemble of five deep learning models, each one created by cascading multiple Inception modules. The Inception Network is inspired by the Inception-v4 architecture Szegedy et al. (2017) and contains two residual blocks in which each residual block's input is transferred via a shortcut linear connection to be added to the next block's input. Following these residual blocks, a Global Average Pooling (GAP) layer averages the outputted multivariate time series by the previous layers. The network contains a softmax activation function for label prediction in the last layer.

Previously to the proposal of InceptionTime, Ismail Fawaz et al. (2020) showed that the Residual Network (ResNet) is a strong deep learning baseline for time series, followed by the Fully Convolutional Network (FCN). In this work, we also evaluate the performance of the Time Convolutional Neural Network (Time-CNN) (Zhao et al. 2017).

FCN is a type of neural network architecture designed for semantic segmentation tasks (e.g., image segmentation), where the goal is to label the regions of an image based on its semantic category. The use of FCNs for time-series classification was first proposed by Wang, Yan, and Oates (2017). The key characteristic of FCNs is that they are composed entirely of convolutional layers without fully connected layers at the end, different from a typical Convolutional Neural Network (CNN). In the context of time series, a convolution consists of sliding one-dimensional filters over the series, enabling the extraction of non-linear discriminant features. Precisely, the FCN architecture proposed by Wang, Yan, and Oates consists of three convolutional blocks with the filter sizes {128, 256, 128}, where each block contains three operations: a convolution followed by batch normalization, and the result is then fed into a Rectified Linear Unit (ReLU) activation function. The convolution operation is fulfilled by three one-dimensional kernels with the sizes {8, 5, 3} without striding. The output features of the third convolutional block are fed into a Global Average Pooling (GAP) layer. Finally, the prediction is produced by a softmax function.

ResNet increases the depth of conventional neural networks by introducing residual shortcut connections between consecutive convolutional layers (Wang, Yan & Oates 2017). Figure 8 illustrates the ResNet architecture evaluated in this work (Ismail Fawaz et al. 2019). The network contains 11 layers, of which the first nine are convolutional layers, followed by a GAP layer that averages the time series across the time dimension. The network has three residual blocks followed by a GAP layer and a final softmax classifier with the number of neurons equals the number of classes. Each residual block comprises three convolutions whose output is

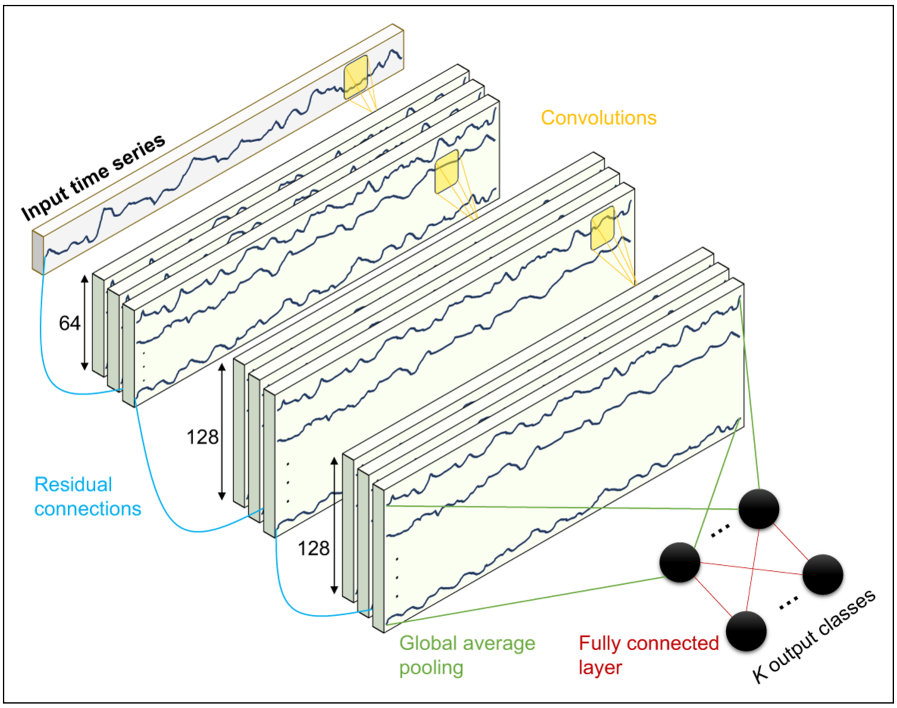


Figure 8 Residual Network (ResNet) architecture for time-series classification (Lima & Souza 2023).

added to the residual block's input and then fed to the next layer. The number of filters for all convolutions is 64, with the ReLU activation function. In each residual block, the filter's length is set to 8, 5, and 3 for the first, second, and third convolution.

A standard CNN architecture consists of two main components: *i)* a feature extraction module composed of convolution layers followed by pooling operations to reduce the dimension of feature maps, and *ii)* a fully connected network that receives the previously extracted features as input. In the context of time-series classification, the Time-CNN architecture comprises two consecutive convolutional layers with 6 and 12 filters, respectively, followed by a local average pooling operation of length 3. In this network, the convolutional layers use the sigmoid activation function. Time-CNN diverges from the typical CNN in two aspects: firstly, it employs the mean squared error (MSE) as the loss function instead of the traditional categorical cross-entropy, and secondly, the network incorporates a local average pooling operation instead of local max pooling.

4. EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

4.1 SETTINGS

We performed the tasks of species and age prediction of malarial mosquitoes using the mid-infrared spectroscopy data provided by Siria et al. (2022). We split the dataset into training and testing sets following the 70/30 ratio. For both tasks, we use classification accuracy to measure the performance of our models. This measure represents the ratio of correct predictions to the total number of examples in the testing set.

Our evaluation considers 14 machine learning algorithms from four time-series learning approaches: *i)* feature-based, *ii)* interval-based, *iii)* convolution-based, and *iv)* deep learning-based. In Table 4.1, we show the set of algorithms from each approach and their parameter values. For feature-based methods, we found the values by a grid search procedure performed

APPROACH	ALGORITHM	PARAMETERS
Feature-based	K-Nearest Neighbors (KNN)	k = 1, Distance: Manhattan
	Logistic Regression (LR)	C = 5, Penalty: L1, Solver: linear
	Support Vector Machines (SVM)	C = 5, Kernel: linear
	Random Forest (RF)	Estimators: 300, Criterion: entropy
	XGBoost (XGB)	Estimators: 300, Learning rate: 0.1, Gamma: 0.1, Max. depth: 7
Interval-based	Time Series Forest (TSF)	Estimators: 200, Intervals: $\sqrt{m}$
	Canonical Interval Forest (CIF)	Estimators: 200, Intervals: $\sqrt{m}$
	Diverse Representation CIF (DrCIF)	Estimators: 200, Intervals: $\sqrt{m}$
Convolution-based	Random Convolutional Kernel Transform (Rocket)	Kernels: 10000
	Minimally Rocket (MiniRocket)	Kernels: 10000, Max. dilations per kernel: 32, Features per kernel: 4
Deep learning-based	Residual Network (ResNet)	Residual blocks: 3, Conv. per residual block: 3, Filters: [128,64,64], Kernel size: [8,5,3], Padding: same, Activation: ReLU, Epochs: 2000
	InceptionTime	Classifiers: 5, Depth: 6, Filters: 32, Conv. per layer: 3, Kernel size: 40, Padding: same, Activation: ReLU, Epochs: 1500
	Fully Convolutional Network (FCN)	Layers: 3, Kernel size: [8,5,3], Filters: [128,256,128], Avg. pool size: 3, Padding: same, Activation: ReLU, Epochs: 2000
	Time Convolutional Neural Network (Time-CNN)	Layers: 2, Kernel size: 7, Filters: [6,12], Avg. pool size: 3, Padding: valid, Activation: sigmoid, Epochs: 2000

in the training data (Syarif, Prugel-Bennett & Wills 2016). For the remaining algorithms, we consider default values.

Since algorithms such as TSF, CIF, DrCif, Rocket, and MiniRocket are unsuitable for non-time-series data, we categorize these algorithms as *non-traditional* along the paper. On the other hand, algorithms such as KNN, LR, SVM, RF, and XGB used in the feature-based approach are categorized as *traditional* since they can be employed on conventional and time-series data.

For approaches such as feature-based, we trained five algorithms considering the following feature sets: *i)* raw data, in which we consider the 1,000 observations of a time series after the preprocessing phase as features, *ii)* wavenumbers, in which we extract the time-series values at 17 representative positions, *iii)* Catch-22, in which we extracted 22 predictive features and *iv)* Catch-22 + wavenumbers, in which we concatenate the 22 features extracted from Catch-22 and 17 wavenumbers, composing a feature vector with 39 features. The second option for feature sets (wavenumbers) is the same as considered by the related works previously discussed. The other feature sets are our attempts to improve the results from the literature.

Given the 14 supervised learning algorithms and feature sets, we have evaluated 29 different settings. Figure 9 shows an overview of our experimental setting considering varying approaches, feature sets, and supervised learning algorithms.

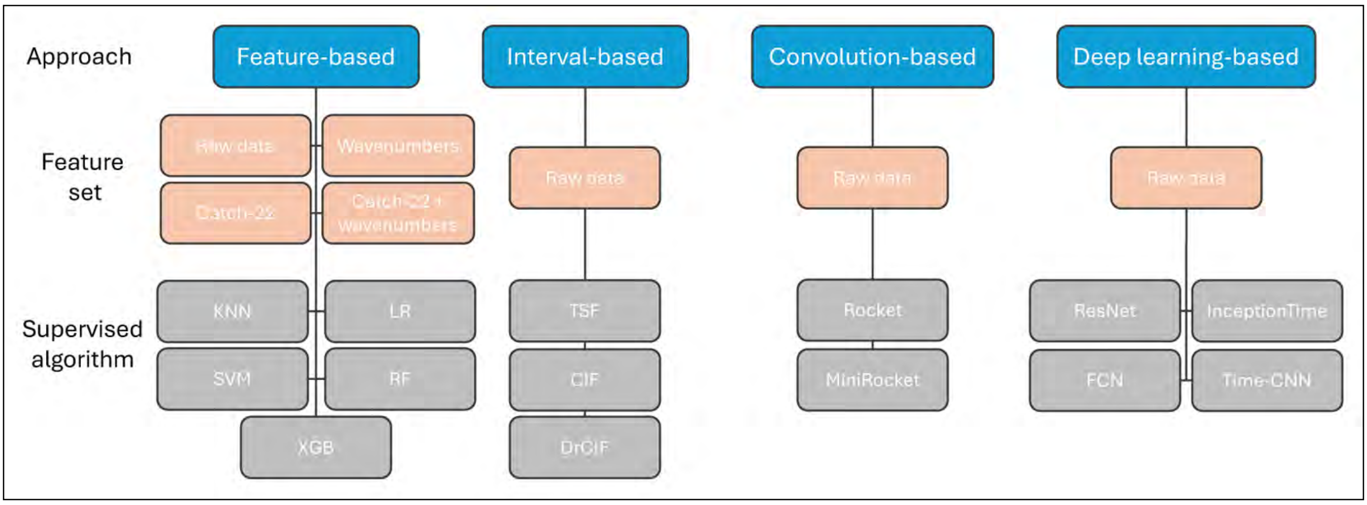


Figure 9 Overview of the varying approaches, feature sets, and supervised learning algorithms covered in the experimental evaluation.

4.2 SPECIES PREDICTION

The species prediction task evaluated in this work aims to distinguish two species from the *Anopheles* genus. The *Anopheles gambiae* and *Anopheles arabiensis* are morphologically indistinguishable in the adult stage (Zianni et al. 2013), and both are the most broadly distributed and efficient vectors of malaria (Coetzee, Craig & Le Sueur 2000).

We begin our analysis by showing the results of the feature-based algorithms. In Figure 10, we show the accuracies obtained by the five algorithms evaluated considering four different feature sets. For all machine learning algorithms, we note that the best feature set for this task is composed of the 1,000 observations of the time series (i.e., raw data), outperforming the wavenumbers as proposed by González Jiménez et al. (2019) in their evaluation with similar

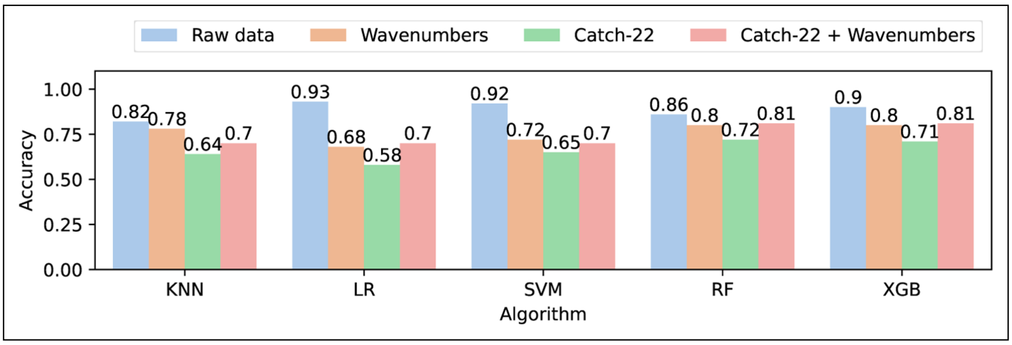
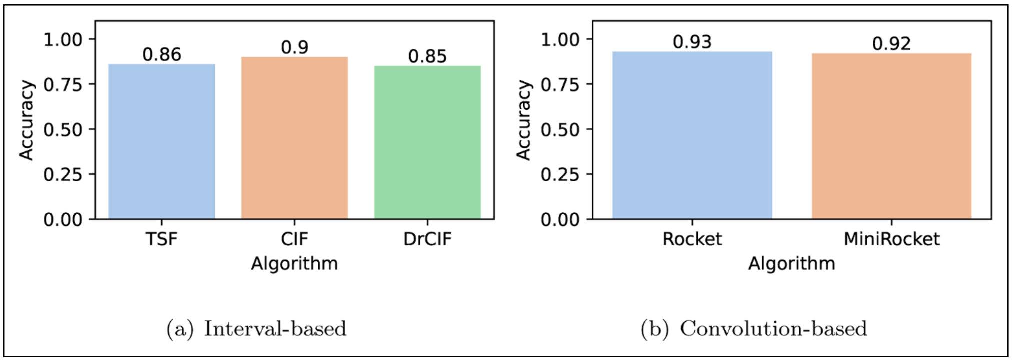


Figure 10 Accuracy of feature-based classifiers for the species prediction.

MIRS data. Using raw data, Logistic Regression (LR) and SVM are the most accurate classifiers, with 93% and 92% of correct predictions, respectively. Besides the performance, it is essential to note an advantage of this feature set: the lack of an additional step for feature extraction as required by other methods. On the other hand, the use of features provided by Catch-22 obtained the worst results, varying from 58% (LR) to 72% (RF) of accuracy.

The feature-based is a well-known approach that allows the use of traditional machine learning algorithms. However, interval and convolution-based classifiers are well-suited approaches for time-series data that still need to be evaluated in the mosquito prediction, as performed in this work. In Figure 11, we show the results obtained by both approaches. For the interval-based methods, CIF achieved the best result with 90% of accuracy, being surpassed by the best results obtained by feature-based algorithms, while the convolution-based method Rocket achieved a competitive result of 93%.



(a) Interval-based (b) Convolution-based

Figure 11 Accuracy results of interval and convolution-based classifiers for the species prediction.

Finally, the results of the fourth approach are shown in Figure 12. InceptionTime was the most accurate with 97%, followed by ResNet with 96%. Such algorithms presented the best results among all approaches. On the other hand, Time-CNN is outperformed by simple algorithms such as Rocket and Logistic Regression, requiring more time and costly hardware (i.e., GPU) for model training.

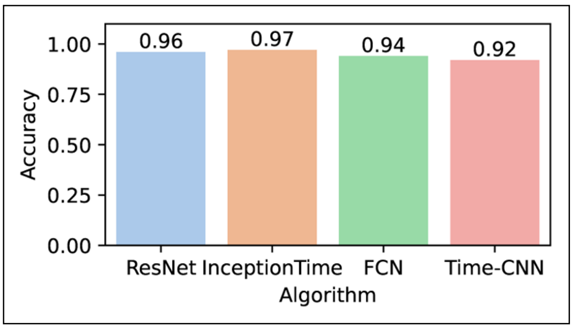


Figure 12 Accuracy of deep learning classifiers for the species prediction.

Table 1 shows a ranking with the 15 best results for better comparison between the approaches, in which we note deep learning and convolutional based methods at the top.

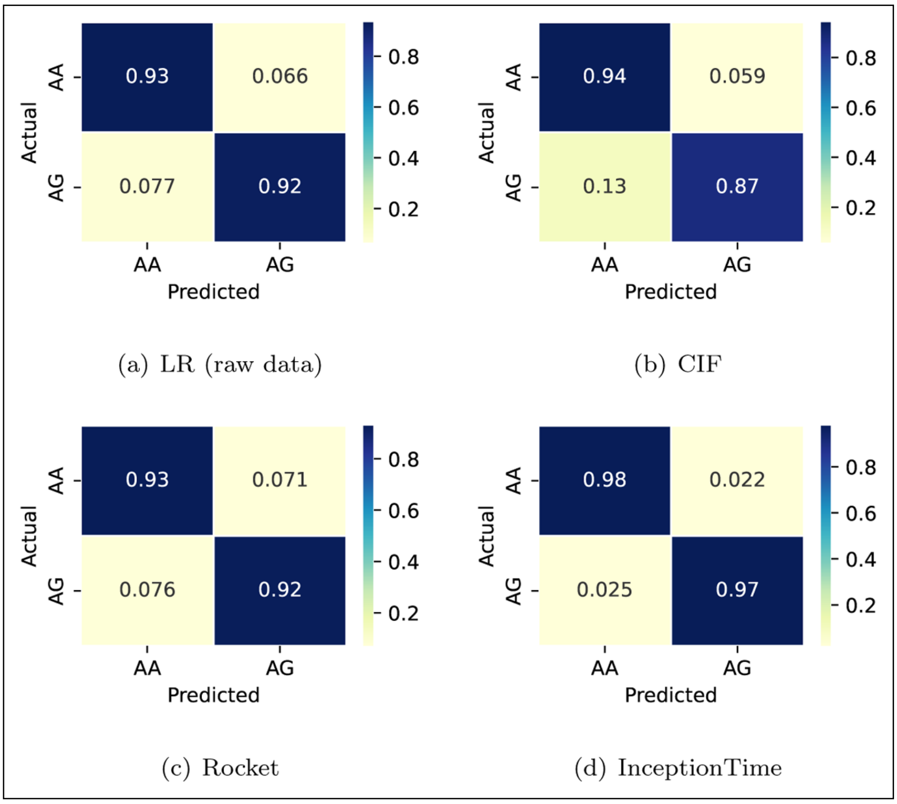
In Figure 13, we show the error distribution between the classes AA (*Anopheles arabiensis*) and AG (*Anopheles gambiae*) carried out by the best algorithm from each approach. The normalized confusion matrix of InceptionTime illustrated in Figure 13d) demonstrates the superior results of this algorithm with predictions close to perfect for both species.

4.3. AGE PREDICTION

The ability of female mosquitoes to transmit diseases such as malaria or yellow fever is age dependent. Due to the incubation period of the parasites and pathogens that mosquitoes transmit, only older mosquitoes are potential vectors of diseases (Dowell, Noutcha & Michel 2011). Thus, predicting the age of malarial mosquitoes holds significant importance in the ongoing efforts to control mosquitoes. Besides, younger mosquitoes are often more susceptible to insecticides, making them prime targets for control interventions. For age prediction, we train our models to distinguish the mosquitos' age using mid-infrared spectroscopy data in three ranges of values: 1–4, 5–10, and 11–17 days.

Table 1 Ranking of algorithms from different categories for the task of species prediction.

ALGORITHM	APPROACH	ACCURACY
InceptionTime	Deep learning	0.97
ResNet	Deep learning	0.96
FCN	Deep learning	0.94
Rocket	Convolution-based	0.93
LR (raw data)	Feature-based	0.93
MiniRocket	Convolution-based	0.92
Time-CNN	Deep learning	0.92
SVM (raw data)	Feature-based	0.92
CIF	Interval-based	0.90
XGB (raw data)	Feature-based	0.90
TSF	Interval-based	0.86
RF (raw data)	Feature-based	0.86
DrCIF	Interval-based	0.85
KNN (raw data)	Feature-based	0.82
RF (Catch-22 + wavenumbers)	Feature-based	0.81



(a) LR (raw data)

(b) CIF

(c) Rocket

(d) InceptionTime

Figure 13 Confusion matrix obtained by the best classifier of each approach (i.e., feature-based, interval-based, convolution-based, and deep learning) for species classification.

In Figure 14, we show the results obtained by feature-based algorithms. As well as for species prediction, using raw data as a feature set provided the best results. However, the most accurate classifier for species prediction (Logistic Regression) did not prove adequate for age prediction. Specifically, the XGBoost algorithm achieved the best result with 75% accuracy, followed by Random Forest (72%), while Logistic Regression shows results close to 50%. Comparatively, we note that age prediction is a more challenging task than species identification.

Figure 15 illustrates the results of interval and convolution-based methods. For age prediction, we noted that the interval-based method CIF slightly outperformed the results of rivals with 76% accuracy.

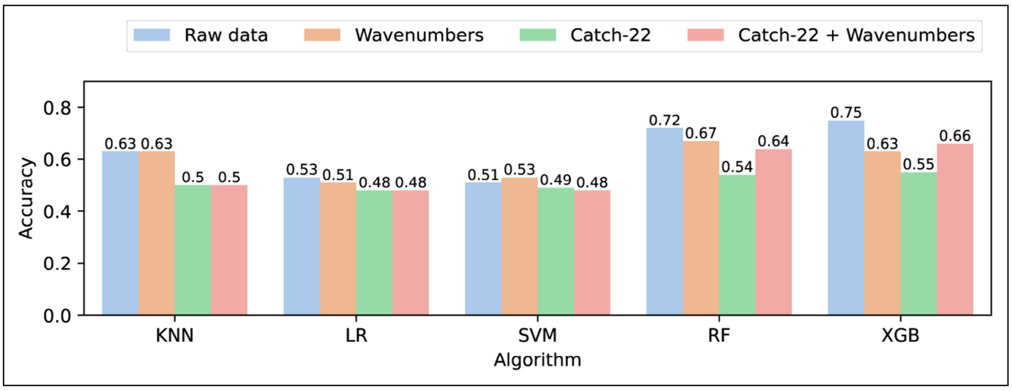
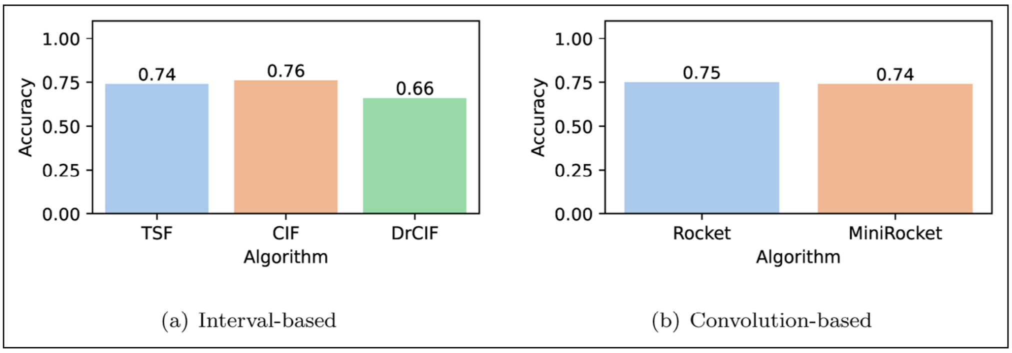


Figure 14 Accuracy of feature-based classifiers for the age prediction.



(a) Interval-based

(b) Convolution-based

Figure 15 Accuracy results of interval and convolution-based classifiers for the age prediction.

As well as for species prediction, the deep learning algorithm InceptionTime achieved the best result among the approaches, achieving 83% accuracy. Figure 16 shows a performance comparison among the deep learning methods. Interestingly, although InceptionTime achieved the best result, the other deep learning algorithms showed results inferior to interval and convolution-based methods.

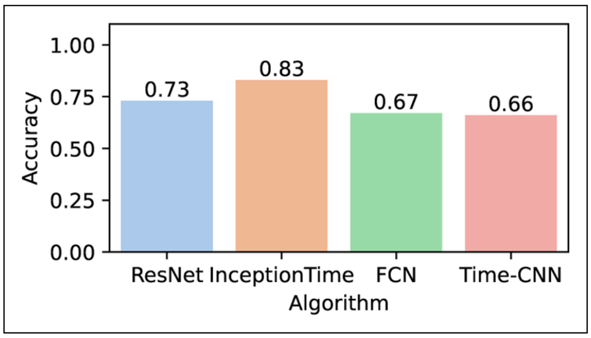


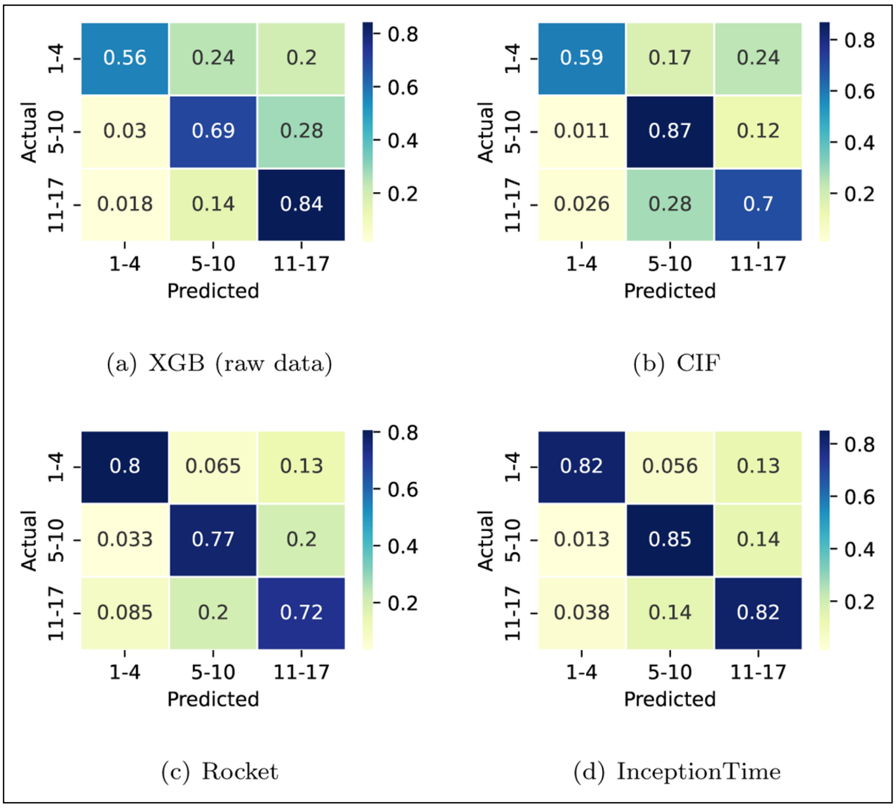
Figure 16 Accuracy of deep learning classifiers for the age prediction.

In the ranking shown in Table 2 with the best 15 algorithms for age prediction, we note only InceptionTime as a deep learning approach at the top, followed by algorithms from the interval, convolution, and feature-based approaches.

In Figure 17, we show the normalized confusion matrix of the best algorithm from each approach. In these results, we can note that although XGB (raw data) and CIF achieved similar accuracies of 75% and 76%, respectively, the algorithms make different mistakes. For example, both algorithms concentrate most of their errors in predicting the age of younger insects (i.e., 1–4 days). However, XGB makes more wrong predictions for insects comprising 5–10 days, while CIF makes mistakes for the 11–17 days range. For InceptionTime, the errors vary from 0.15 to 0.18 for all age grades, being an adequate algorithm for age prediction.

ALGORITHM	APPROACH	ACCURACY
InceptionTime	Deep learning	0.83
CIF	Interval-based	0.76
Rocket	Convolution-based	0.75
XGB (raw data)	Feature-based	0.75
MiniRocket	Convolution-based	0.74
TSF	Interval-based	0.74
ResNet	Deep learning	0.73
RF (raw data)	Feature-based	0.72
FCN	Deep learning	0.67
RF (Wavenumbers)	Feature-based	0.67
Time-CNN	Deep learning	0.66
DrCIF	Interval-based	0.66
XGB (Catch-22 + wavenumbers)	Feature-based	0.66
RF (Catch-22 + wavenumbers)	Feature-based	0.64
KNN (raw data)	Feature-based	0.63

Table 2 Ranking of algorithms from different categories for the task of age prediction.



(a) XGB (raw data)

(b) CIF

(c) Rocket

(d) InceptionTime

Figure 17 Confusion matrix obtained by the best classifier of each approach (i.e., feature-based, interval-based, convolution-based, and deep learning-based) for species classification.

5. DISCUSSION

In the previous section, we analyzed the results of algorithms from different approaches. From this analysis, we can conclude that time series mining classifiers such as InceptionTime, Rocket, and CIF outperformed the traditional machine learning approach previously evaluated in the literature, such as an SVM model induced with specific wavenumbers from the spectroscopy data as features. However, it is essential to understand the limitations of each approach for their employment in practice. Therefore, we discuss the advantages and limitations of each approach to guide this choice.

FEATURE-BASED APPROACH

The main advantages of this approach are the interpretability and flexibility. In some problems, the time-series shape of raw data can be well-defined according to the class (e.g., with peaks and valleys located in specific regions), helping to visually understand the classification result provided by a similarity-based algorithm as the nearest neighbor. However, the difference between the shapes of time series from different species is subtle for MIRS data, as illustrated in Figure 3. In this case, analyzing specific wavenumbers or the information extracted by the Catch-22 can support understanding the data and results. Regarding flexibility, the features can be used to train virtually any machine learning algorithm, such as Naive Bayes, Decision Trees, and Artificial Neural Networks. The limitation of this approach is that its performance heavily relies on selecting and engineering relevant features, which can be time-consuming and dependent on a domain expert. Moreover, most cases involve manually executing an additional step in the learning process. For example, the selection of 17 wavenumbers provided by González Jiménez et al. (2019) was performed through the careful analysis of chemistry and ecologist experts.

INTERVAL-BASED APPROACH

Such an approach is robust to noise and outliers since it discards most observations after selecting relevant and minor subsequences from the raw data. Once the intervals are defined, any machine learning algorithm can be trained using them, bringing flexibility to the approach. Optionally, we can extract features from the intervals to train a more robust classifier, as performed by CIF and DrCIF. An essential advantage of this approach is its efficiency and scalability in processing high-dimensional data, a usual characteristic of time series. A limitation of the interval-based approach is the loss of potentially relevant information during the interval selection phase and the influence of interval size, a parameter that requires careful attention.

CONVOLUTION-BASED APPROACH

Rocket achieves comparable accuracy to the state-of-the-art ensemble methods composed of three dozen classifiers, such as HIVE-COTE (Lines, Taylor & Bagnall 2018) or complex deep learning methods, with a fraction of processing time. However, since the algorithm uses 10,000 random convolutional kernels in terms of length, weights, bias, dilation, and padding to capture relevant features, it is a non-deterministic solution, which could be considered a limitation of this approach regarding reproducibility and predictability. However, we show that MiniRocket provided similar results with a deterministic solution. We also point out that a limitation of this approach is the requirement for a lengthy time series, which is not an issue for MIRS data.

DEEP LEARNING-BASED APPROACH

The classification accuracy is the main advantage of the deep learning-based approach, mainly for InceptionTime, an algorithm adapted for time series data. However, deep learning algorithms such as FCN and Time-CNN performed inferiorly than simple feature-based algorithms. The unified framework of deep learning algorithms that perform both feature extraction and classification steps is advantageous, removing the need for manual feature engineering. This approach's limitations are the need for a large amount of labeled data for training, lack of model interpretability, and time costs for model training. While models from alternative approaches were trained using CPU processing, the algorithms employed in this methodology necessitated GPU utilization for model training within an acceptable timeframe.

6. CONCLUSIONS

Estimating and monitoring mosquito species' population and age are essential for assertive employment of control measures, such as using adequate adulticides and larvicides in a potential risk region ahead of outbreaks. Besides, this monitoring helps evaluate the effectiveness of current control methods. The current methods for accurate monitoring are costly and time-consuming, such as those based on molecular analysis as PCR. In the last years, researchers have been investigating rapid and low-cost alternatives. Mid-infrared spectroscopy is the most recent and prominent technique, which generates time series data.

The species and age identification using MIRS data depends on machine learning algorithms to recognize the complex relationships in the spectrum. Previous studies consider traditional machine learning algorithms that do not consider the particularities of time series data, such as temporal dependencies and correlations between features. This work comprehensively evaluates the state-of-the-art classification methods for time series data employed in the public health application of mosquito species identification and age prediction.

Our experiments considering algorithms from different time series mining approaches show that the deep learning algorithm InceptionTime is the most accurate method, able to identify 97% of species correctly and predict the age of insects with 83% accuracy, outperforming the current results from the literature using traditional machine learning algorithms. Compared with complex deep learning methods, the convolution-based algorithm Rocket presents competitive results with a low computational cost. The results obtained by these algorithms outperform the state-of-the-art, which considers feature-based methods and convolutional neural networks. Thus, this research contributes to the field by highlighting the effectiveness of time series mining approaches for malaria vector control using spectroscopy.

DATA ACCESSIBILITY STATEMENT

Data supporting this study are openly available from Enlighten database and are available at https://doi.org/10.5525/gla.researchdata.1235.

ACKNOWLEDGEMENTS

The authors thank Google for the grant in the Google Cloud research credits program that supported this research.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

LC, HC, and VS conceived this research and designed experiments; LC preprocessed and cleaned the data; HC and VS performed experiments and analysis; LC and HC collaborated to interpret the results. LC and VS drafted the manuscript. All authors read and approved the final manuscript.

AUTHOR AFFILIATIONS

Lucas G. M. Castro orcid.org/0009-0007-5717-4705
Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição 1155, Curitiba, PR, Brazil

Henrique V. Costa
Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição 1155, Curitiba, PR, Brazil

Vinicius M. A. Souza orcid.org/0000-0003-3175-7922
Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição 1155, Curitiba, PR, Brazil

REFERENCES

Bagnall, A, Lines, J, Bostrom, A, Large, J and **Keogh, E** 2017 The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery,* 31: 606–660. DOI: https://doi.org/10.1007/s10618-016-0483-9

Barbosa, T M, de Lima, L A, Dos Santos, M C, Vasconcelos, S D, Gama, R A and **Lima, K M** 2018 A novel use of infra-red spectroscopy (nirs and atr-ftir) coupled with variable selection algorithms for the identification of insect species (diptera: Sarcophagidae) of medico-legal relevance. *Acta Tropica,* 185: 1–12. DOI: https://doi.org/10.1016/j.actatropica.2018.04.025

Batista, G, Hao, Y, Keogh, E and **Mafra-Neto, A** 2011 Towards automatic classification on flying insects using inexpensive sensors. In *International Conference on Machine Learning and Applications* (pp. 364–369). IEEE volume 1. DOI: https://doi.org/10.1109/ICMLA.2011.145

Breiman, L 2001 Random forests. *Machine learning,* 45: 5–32. DOI: https://doi.org/10.1023/A:1010933404324

Broersen, P M and **de Waele, S** 2000 Some benefits of aliasing in time series analysis. In *European Signal Processing Conference* (pp. 1–4). IEEE. https://ieeexplore.ieee.org/document/7075299

Caminade, C, Kovats, S, Rocklov, J, Tompkins, A M, Morse, A P, Colón-González, F J, Stenlund, H, Martens, P and **Lloyd, S J** 2014 Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences,* 111: 3286–3291. DOI: https://doi.org/10.1073/pnas.1302089111

Cella, W, Baia-da Silva, D C, Melo, G C D, Tadei, W P, Sampaio, V D S, Pimenta, P, Lacerda, M V G and **Monteiro, W M** 2019 Do climate changes alter the distribution and transmission of malaria? Evidence assessment and recommendations for future studies. *Revista da Sociedade Brasileira de Medicina Tropical,* 52: e20190308. DOI: https://doi.org/10.1590/0037-8682-0308-2019

Coetzee, M, Craig, M and **Le Sueur, D** 2000 Distribution of African malaria mosquitoes belonging to the anopheles gambiae complex. *Parasitology Today,* 16: 74–77. DOI: https://doi.org/10.1016/S0169-4758(99)01563-X

Dempster, A, Petitjean, F and **Webb, G I** 2020 Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery,* 34: 1454–1495. DOI: https://doi.org/10.1007/s10618-020-00701-z

Dempster, A, Schmidt, D F and **Webb, G I** 2021 Minirocket: A very fast (almost) deterministic transform for time series classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 248–257). DOI: https://doi.org/10.1145/3447548.3467231

Dempster, A, Schmidt, D F and **Webb, G I** 2023 Hydra: Competing convolutional kernels for fast and accurate time series classification. *Data Mining and Knowledge Discovery,* 37(5): 1–27. DOI: https://doi.org/10.1007/s10618-023-00939-3

Deng, H, Runger, G, Tuv, E and **Vladimir, M** 2013 A time series forest for classification and feature extraction. *Information Sciences,* 239: 142–153. DOI: https://doi.org/10.1016/j.ins.2013.02.030

Dowell, F E, Noutcha, A E and **Michel, K** 2011 The effect of preservation methods on predicting mosquito age by near infrared spectroscopy. *The American Journal of Tropical Medicine and Hygiene*, 85: 1093. DOI: https://doi.org/10.4269/ajtmh.2011.11-0438

Fernandes, M S, Cordeiro, W and **Recamonde-Mendoza, M** 2021 Detecting Aedes Aegypti mosquitoes through audio classification with convolutional neural networks. *Computers in Biology and Medicine,* 129: 104152. DOI: https://doi.org/10.1016/j.compbiomed.2020.104152

Fulcher, B D and **Jones, N S** 2017 HCTSA: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems,* 5: 527–531. DOI: https://doi.org/10.1016/j.cels.2017.10.001

González Jiménez, M, Babayan, S A, Khazaeli, P, Doyle, M, Walton, F, Reddy, E, Glew, T, Viana, M, Ranford-Cartwright, L, Niang, A, et al. 2019 Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. *Wellcome Open Research,* 4: 76. DOI: https://doi.org/10.12688/wellcomeopenres.15201.3

Ismail Fawaz, H, Forestier, G, Weber, J, Idoumghar, L and **Muller, P-A** 2019 Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery,* 33: 917–963. DOI: https://doi.org/10.1007/s10618-019-00619-1

Ismail Fawaz, H, Lucas, B, Forestier, G, Pelletier, C, Schmidt, D F, Weber, J, Webb, G I, Idoumghar, L, Muller, P-A and **Petitjean, F** 2020 Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery,* 34: 1936–1962 DOI: https://doi.org/10.1007/s10618-020-00710-y

Johnson, B J, Hugo, L E, Churcher, T S, Ong, O T and **Devine, G J** 2020 Mosquito age grading and vector-control programmes. *Trends in Parasitology,* 36: 39–51. DOI: https://doi.org/10.1016/j.pt.2019.10.011

Johnson, J B 2020 An overview of near-infrared spectroscopy (NIRS) for the detection of insect pests in stored grains. *Journal of Stored Products Research,* 86: 101558. DOI: https://doi.org/10.1016/j.jspr.2019.101558

Kahn, M C, Celestin, W and **Offenhauser, W** 1945 Recording of sounds produced by certain disease-carrying mosquitoes. *Science,* 101: 335–336. DOI: https://doi.org/10.1126/science.101.2622.335

Lawrence, S, Giles, C L, Tsoi, A C and **Back, A D** 1997 Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks,* 8: 98–113. DOI: https://doi.org/10.1109/72.554195

LeCun, Y, Bengio, Y and **Hinton, G** 2015 Deep learning. *Nature,* 521: 436–444. DOI: https://doi.org/10.1038/nature14539

Lima, F T and **Souza, V M A** 2023 A large comparison of normalization methods on time series. *Big Data Research,* 34: 100407. DOI: https://doi.org/10.1016/j.bdr.2023.100407

Lines, J, Taylor, S and **Bagnall, A** 2018 Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data,* 12: 1–35. DOI: https://doi.org/10.1145/3182382

Lubba, C H, Sethi, S S, Knaute, P, Schultz, S R, Fulcher, B D and **Jones, N S** 2019 Catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. *Data Mining and Knowledge Discovery,* 33: 1821–1852. DOI: https://doi.org/10.1007/s10618-019-00647-x

Lubinda, J, Haque, U, Bi, Y, Hamainza, B and **Moore, A J** 2021 Near-term climate change impacts on sub-national malaria transmission. *Scientific Reports,* 11: 751. DOI: https://doi.org/10.1038/s41598-020-80432-9

Main, F 1909 La destruction des fourmis blanches. *Journal d'Agriculture Tropicale,* 101: 350.

Mankin, R W, Hagstrum, D W, Smith, M T, Roda, A and **Kairo, M T** 2011 Perspective and promise: A century of insect acoustic detection and monitoring. *American Entomologist,* 57: 30–44. DOI: https://doi.org/10.1093/ae/57.1.30

Mayagaya, V S, Michel, K, Benedict, M Q, Killeen, G F, Wirtz, R A, Ferguson, H M and **Dowell, F E** 2009 Non-destructive determination of age and species of *Anopheles gambiae* sl using near-infrared spectroscopy. *The American Journal of Tropical Medicine and Hygiene,* 81: 622–630. DOI: https://doi.org/10.4269/ajtmh.2009.09-0192

Middlehurst, M, Large, J and **Bagnall, A** 2020 The canonical interval forest (CIF) classifier for time series classification. In *IEEE International Conference on Big Data* (pp. 188–195). IEEE. DOI: https://doi.org/10.1109/BigData50022.2020.9378424

Middlehurst, M, Large, J, Flynn, M, Lines, J, Bostrom, A and **Bagnall, A** 2021 Hive-cote 2.0: A new meta ensemble for time series classification. *Machine Learning,* 110: 3211–3243. DOI: https://doi.org/10.1007/s10994-021-06057-9

Modu, B, Polovina, N, Lan, Y, Konur, S, Asyhari, A T and **Peng, Y** 2017 Towards a predictive analytics-based intelligent malaria outbreak warning system. *Applied Sciences,* 7: 836. DOI: https://doi.org/10.3390/app7080836

Moore, A 1991 Artificial neural network trained to identify mosquitoes in flight. *Journal of Insect Behavior,* 4: 391–396. DOI: https://doi.org/10.1007/BF01048285

Mwanga, E P, Mapua, S A, Siria, D J, Ngowo, H S, Nangacha, F, Mgando, J, Baldini, F, González Jiménez, M, Ferguson, H M, Wynne, K, et al. 2019 Using mid-infrared spectroscopy and supervised machine-learning to identify vertebrate blood meals in the malaria vector, anopheles arabiensis. *Malaria Journal,* 18: 1–9. DOI: https://doi.org/10.1186/s12936-019-2822-y

Parmezan, A R, Souza, V M, Žliobaitė, I and **Batista, G E** 2021 Changes in the wing-beat frequency of bees and wasps depending on environmental conditions: A study with optical sensors. *Apidologie,* 52: 731–748. DOI: https://doi.org/10.1007/s13592-021-00860-y

Potamitis, I and **Rigakis, I** 2016 Large aperture optoelectronic devices to record and time-stamp insects' wingbeats. *IEEE Sensors Journal,* 16: 6053–6061. DOI: https://doi.org/10.1109/JSEN.2016.2574762

Ritchie, S A, Devine, G J, Vazquez-Prokopec, G M, Lenhart, A E, Manrique-Saide, P and **Scott, T W** 2021 Insecticide-based approaches for dengue vector control. In *Ecology and Control of Vector-Borne Diseases* (pp. 380–390). Wageningen Academic Publishers. DOI: https://doi.org/10.3920/978-90-8686-895-7_4

Santolamazza, F, Mancini, E, Simard, F, Qi, Y, Tu, Z and **della Torre, A** 2008 Insertion polymorphisms of sine200 retrotransposons within speciation islands of anopheles gambiae molecular forms. *Malaria Journal,* 7: 1–10. DOI: https://doi.org/10.1186/1475-2875-7-163

Sarker, I H 2021 Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science,* 2: 420. DOI: https://doi.org/10.1007/s42979-021-00815-1

Silva, D F, Souza, V M A, Ellis, D P W, Keogh, E J and **Batista, G** 2015 Exploring low-cost laser sensors to identify flying insect species: Evaluation of machine learning and signal processing methods. *Journal of Intelligent & Robotic Systems,* 80: 313–330. DOI: https://doi.org/10.1007/s10846-014-0168-9

Siria, D J, Sanou, R, Mitton, J, Mwanga, E P, Niang, A, Sare, I, Johnson, P C, Foster, G M, Belem, A M, Wynne, K, et al. 2022 Rapid age-grading and species identification of natural mosquitoes for malaria surveillance. *Nature Communications,* 13: 1501 DOI: https://doi.org/10.1038/s41467-022-28980-8

Souza, V M A 2017 Identifying Aedes Aegypti mosquitoes by sensors and one-class classifiers. In *Iberoamerican Congress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (pp. 10–18). Springer. DOI: https://doi.org/10.1007/978-3-319-52277-7_2

Souza, V M A, dos Reis, D M, Maletzke, A G and **Batista, G** 2020 Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery,* 34: 1805–1858. DOI: https://doi.org/10.1007/s10618-020-00698-5

Souza, V M A, Silva, D F and **Batista, G E A P A** 2014 Extracting texture features for time series classification. In *International Conference on Pattern Recognition* (pp. 1425–1430). DOI: https://doi.org/10.1109/ICPR.2014.254

Sriwichai, P, Karl, S, Samung, Y, Sumruayphol, S, Kiattibutr, K, Payakkapol, A, Mueller, I, Yan, G, Cui, L and **Sattabongkot, J** 2015 Evaluation of CDC light traps for mosquito surveillance in a malaria endemic area on the Thai-Myanmar border. *Parasites & Vectors,* 8: 1–10. DOI: https://doi.org/10.1186/s13071-015-1225-3

Syarif, I, Prugel-Bennett, A and **Wills, G** 2016 Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *Telecommunication Computing Electronics and Control,* 14: 1502–1509 DOI: https://doi.org/10.12928/telkomnika.v14i4.3956

Szegedy, C, Ioffe, S, Vanhoucke, V and **Alemi, A** 2017 Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 4278–4284). DOI: https://doi.org/10.1609/aaai.v31i1.11231

Wang, Z, Yan, W and **Oates, T** 2017 Time series classification from scratch with deep neural networks: A strong baseline. In *International Joint Conference on Neural Networks* (pp. 1578–1585). DOI: https://doi.org/10.1109/IJCNN.2017.7966039

World Health Organization (WHO) 2023 World Health Organization (WHO): World malaria report 2023 Available at https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023 Last accessed December 2023.

Zhao, B, Lu, H, Chen, S, Liu, J and **Wu, D** 2017 Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics,* 28: 162–169. DOI: https://doi.org/10.21629/JSEE.2017.01.18

Zhao, Z-Q, Zheng, P, Xu, S-t and **Wu, X** 2019 Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems,* 30: 3212–3232. DOI: https://doi.org/10.1109/TNNLS.2018.2876865

Zianni, M R, Nikbakhtzadeh, M R, Jackson, B T, Panescu, J and **Foster, W A** 2013 Rapid discrimination between anopheles gambiae ss and anopheles arabiensis by high-resolution melt (HRM) analysis. *Journal of Biomolecular Techniques,* 24: 1. DOI: https://doi.org/10.7171/jbt.13-2401-001

TO CITE THIS ARTICLE:
Castro, LGM, Costa, HV and Souza, VMA 2024 Time-Series Mining Approaches for Malaria Vector Prediction On Mid-Infrared Spectroscopy Data. *Data Science Journal*, 23: 25, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2024-025

Submitted: 17 February 2024
Accepted: 13 April 2024
Published: 01 May 2024