Detection of Zika virus infection on mosquitoes using spectroscopy and machine learning

L. M. Reigoto[†], R. M. de Freitas[§], G. M. Araujo[‡†], A. A. de Lima[⋆],

***Abstract*— This work shows a method to classify mosquitoes infected with the Zika virus. We accomplish that by using spectroscopy and machine learning. Our model takes the light absorbance of wavelengths from 350 to 1000 nm as inputs. It employs a combination of Linear Discriminant Analysis (LDA) of the windowed version of the signal (to take advantage of nonlinearities) and Support Vectors Machine (SVM) to classify the samples. The proposed method can detect the presence of the Zika virus with 100% accuracy in less than 7 days post-infection. The accuracy drops to 77.7% when 10 days have passed. The main advantages are the low cost and the possibility to make predictions in real-time.**

***Keywords*— Zika virus, Machine Learning, Detection, Arboviruses, Classification, Support Vector Machine.**

I. Introduction

The Zika virus can cause serious side effects like microcephaly, congenital Zikavirus syndrome, and Guillain-Barré syndrome. In 2015 this disease started an epidemic in America. In February 2016, the World Health Organization declared a public health emergency due to the microcephaly cases associated with the Zika virus epidemic. The traditional method to detect this virus is using RT-qPCR [10], [4], [11]. This method takes a considerable time, is expensive, intrusive, and requires skilled workers. The method described in [6] employs near-infrared spectroscopy (NIRS) with wavelengths from 700 to 2500 nm to detect Zika virus in mosquitoes. The spectra were mean-centered and then classified using the Partial Least Squares (PLS) regression method in GRAMS Plus/ IQ software (Thermo Galactic). Table I shows the results.

NIRS method						
Dataset	4 DPI (%)		7 DPI (%)		10 DPI (%)	
	TPR	SPC	TPR	SPC	TPR	SPC
Cohort 1 train	83.3	96.8	93.5	96.4	-	-
Cohort 1 Validation (sorted)	100.0	94.1	100.0	100.0	-	-
Cohort 2 head/thorace	98.7	98.3	100.0	98.3	100.0	86.7
Cohort 2 abdomen	98.7	85.0	96.2	80.0	97.4	68.3

TABLE I: NIRS method [6]

The main problem of the method described in [6] is its cost. A near-infrared spectrograph can cost thousands of dollars. In this work, we employ machine learning methods to verify if it is possible to detect the Zika virus in mosquitoes using a narrower wavelength band. To do so, we employ Linear Discriminant Analysis (LDA) and Support Vectors Machine (SVM) on the same data from [6] but considering only wavelengths from 350 to 1000 nm. In terms of cost, one can find a full near-infrared (NIR) spectrometer (from 350 to 2500 nm) with prices ranging from US $\$95,000$ to US $\$115,000$[1]. However, some spectrometers detect light with wavelengths from 350 to 1000 nm under US $\$1,000$.

This work is organized as follows. Section II describes the dataset introduced in [6] and used here. Section III has the proposed method, the results, and a discussion about them. Conclusions are in Section IV.

II. Dataset

The dataset used in this study was obtained from a partnership with FioCruz researchers. The data was collected by *Laboratório de Mosquitos Transmissores de Hematozoários, Pavilhão Carlos Chagas, Instituto Oswaldo Cruz*, Rio de Janeiro, Brazil. It consists of three pieces of information.

- The absorbance of the measured wavelenghts (from 350 to 2500 nm);
- If the mosquito is infected or not;
- How many Days Post Infection (DPI).

Figure 1 shows the absorbance versus wavelength of two samples. The blue line corresponds to an infected mosquito (7 DPI), and the red line corresponds to an uninfected mosquito.

On this experiment they used *Aedes aegypti* mosquitoes with 5 and 6 days old. Some were feed with blood contaminated with Zika virus, and some were with healthy blood. We had two different cohorts of mosquitoes, each collected and measured independently. The measures are from mosquitos after 4, 7, and 10 Days Post Infection (DPI). The mosquitoes were killed right before the readings by placing them in a closed jar with an acetate-soaked cotton ball for 1 min. After that a NIRS spectrometer was used to measure the absorbance

[‡]Program of Electrical Engineering, Federal Centre of Technological Education of Rio de Janeiro (CEFET/RJ), Campus Maracanã, Rio de Janeiro, RJ, Brazil. E-mail: gabriel.araujo@cefet-rj.br;
[†]Dept. of Automation and Control Engineering, Federal Centre of Technological Education of Rio de Janeiro (CEFET/RJ), Campus Nova Iguaçu, Nova Iguaçu, RJ, Brazil. E-mail: leonardo.reigoto@aluno.cefet-rj.br;
[⋆]Dept. of Telecommunications, Federal Centre of Technological Education of Rio de Janeiro (CEFET/RJ), Campus Nova Iguaçu, Nova Iguaçu, RJ, Brazil. E-mail: amaro.lima,@cefet-rj.br.
[§]Instituto Oswaldo Cruz, Laboratório de Mosquitos Transmissores de Hematozoários, Rio de Janeiro, Rio de Janeiro, Brazil. Email: macieldefreitas@gmail.com

[1]models searched were ATP9110-25H and ATP9110-25H on optosky.com

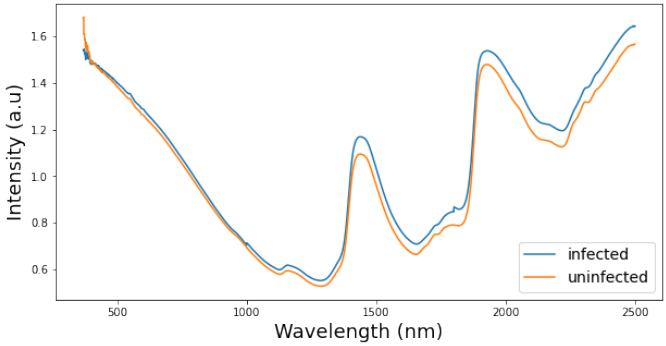

Fig. 1: Example of the difference between the average spectrum of the mosquitoes infected and non-infected. This graph plots the result of the spectroscopy (intensity of light) versus the measured wavelenghts (350 to 2500 nm).

of wavelengths from 350 to 2500 nm. The spectrometer used was a LabSpec 4 i NIR spectrometer from Malvern Panalytical with an internal 18.6-W light source.

The mosquitoes of cohort 1 were measured in their head and thoraces. The mosquitos of cohort 2 were measured in their head, thoraces and abdomen. After the measures, a RT-qPCR test confirms if the mosquito is infected or not.

The samples of each cohort were separated based on their DPI and status (infected or not infected). This left us with up to 6 groups under each cohort as showed on Figures 2 and 3.

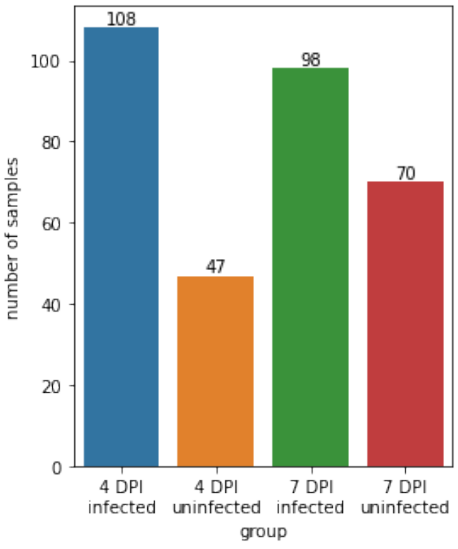

Fig. 2: Distribution of samples from cohort 1.

The data from Cohort 1 contains:

- 4 DPI : 108 infected samples and 47 uninfected samples.
- 7 DPI: 98 infected samples and 70 uninfected samples.

The data from Cohort 2 contains:

- 4 DPI : 76 infected samples and 59 uninfected samples.
- 7 DPI: 77 infected samples and 59 uninfected samples.
- 10 DPI: 77 infected samples and 60 uninfected samples.

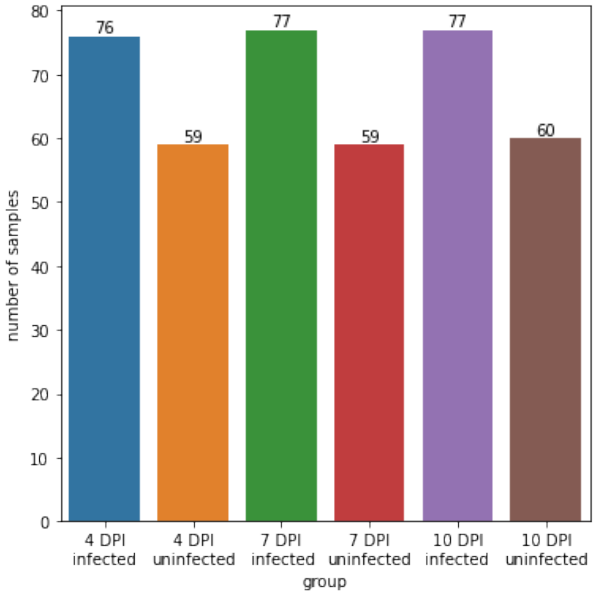

Fig. 3: Distribution of samples from cohort 2.

The spectrometer used combines three devices:

- A silicon sensor that operates between 350 and 1000 nm;
- An InGaAs sensor that operates between 1001 and 1800 nm;
- An InGaAs sensor that operates between 1801 and 2500 nm.

Our model used only the readings from the first device (the silicon sensor).

III. Methodology and Results

In this paper, we use samples from cohort 1 to train our model and test it with data from cohort 2. By doing so, we have the advantage that training and testing data were collected and measured independently. Since the process of preparation of the mosquitoes and measurements can be very complex, this data separation helps ensure that our model can generalize well and predict the status of new samples instead of fitting on noises in the data.

There are two problems related to the data remaining. Its low amount of samples and high dimensionality. Although excluding samples from cohort 2 from our training data increases the problem caused by our low amount of data, the advantages outweigh the disadvantages. It is critical to ensure that our model can generalize well. We used the K-fold method (K = 10) to validate our model using samples from cohort 1 without needing to remove them from our training samples. It helps prevent a higher bias caused by sorting a few samples from a low amount of data.

The K-fold method consists in distributing the data between K different folds. Then we train our data K times, each time removing one folder from the training data and using it to evaluate the model. The final evaluating metrics are the average from all the K models. The final model is the combination of all the models. For classification problems, the class can be the most voted among all models. It is common

to evaluate a model with K-fold and then get a new model using all the folds for training. If we get a reasonable amount of data and folds to reduce noises from specific samples, it is expected by the theory of generalization [1] that our final model will be close to the combination of the K original ones.

A. Applying LDA

One can solve the dimensionality problem by discarding some data with little information. Since our data is composed of absorbance versus wavelength and we are trying to classify it using a continuous interval, its expected some degree of linearity. It is possible to use linearity to deal with dimensionality problems using Linear Discriminant Analysis (LDA) [8], [7] as a feature extractor. It transforms the parameter axes into new ones that best discriminate our classes. The main focus of the LDA is to find a projection axis that maximizes the variance between classes while minimizing the inner variance of each class. The transformed data can discriminate between classes with fewer parameters, so one can discard some of them.A comparison between the used spectral band (350 to 1000 nm) of infected and non-infected mosquitoes can be seen in Fig. 4.

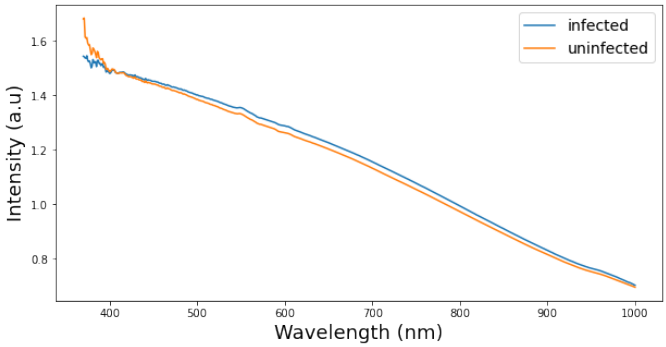

Fig. 4: Example of the difference between the average spectrum of the mosquitoes infected and non-infected. This graph plots the result of the spectroscopy (intensity of light) versus the measured wavelenghts (350 to 1000 nm).

We use the LDA algorithm to extract new features from our data. When used in classification problems with C classes, the LDA algorithm returns $C-1$ axis that best discriminates the classes. Since the problem proposed here only has 2 classes (infected or not), the LDA algorithm returns only 1 axis. Then our model would only be able to find a threshold point on this axis and classify it based on a threshold. The problem with this approach is that it does not deal with non-linearity which could increase the algorithm power. To deal with this we selected a reduced wavelength in the range of $(350, 1000]$ nm, corresponding to the i^{th} intensity vector $\mathbf{x}_i = [x(1), \ldots, x(650)]$ and divided it into twenty-six non-overlapping windows with intervals of 25 wavelengths with ranges of $(350, 375]$, ..., $(975, 1000]$ nm. We generated LDA's for each window associated data, $\text{LDA}_1, \ldots, \text{LDA}_{26}$, and applied the vector window to its respective LDA coordinates, producing the projection of the vector window in LDA space. The twenty-six projections are concatenated creating a new feature vector, $[x'_1, \ldots, x'_{26}]$. This vector is composed of an axis that best discriminates each vector window. Each axis still could be used to classify each window by a threshold, but they can also represent the reliability of these classifications by the distance of the threshold.

After having the feature vector obtained by concatenating the LDA's projections, we use the Support Vector Machine (SVM) to classify the data to obtain the output y_i associated to $\mathbf{x}_i$. This approach was named LDA+SVM method 1, and its block diagram is in Fig. 5.

Another approach, called LDA+SVM method 2, also evaluated in this work, consisted in adding a new feature, x'_{all}, to the previously designed feature vector, making a 27-dimensionality vector. This new feature is the projection of $\mathbf{x}_i$ in the LDA coordinates of the entire interval, LDA_{all}. After this, we apply the SVM classifier. The whole system is in Fig. 5.

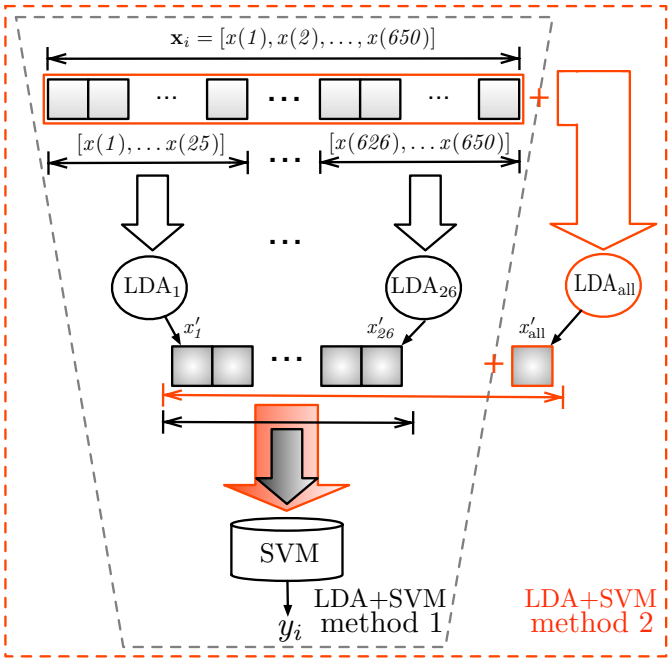

Fig. 5: Block diagram of LDA+SVM method 1 and LDA+SVM method 2.

B. SVM for classification

Support Vector Machine (SVM) [9], [2], [5], [3] is a very powerful and popular algorithm. It can reproduce results of higher complexity models using lower complexity. It makes this algorithm able to handle more parameters using fewer data and with more resistance to the *curse of dimensionality* [2] and *overfitting*.

The idea behind SVM is to use the hyperspace of the inputs parameters and choose the best hyperplane that splits two classes. This hyperplane maximizes the margin between the classes which is calculated by the support vectors. Support vectors are the data points that touch the margin of the hyperplane. This margin classifier can also deal with non-linearly separable classes transforming the data to a higher-dimensional space *phi*, which in practice is implemented by

kernel functions, where the most common ones are kernel polynomial and radial basis functions.

We feed our new vector generated by the LDA's algorithms as an input for a Support Vector Machine (SVM) algorithm with a polynomial kernel. Since we are using a non-linear kernel, we are adding non-linearity to our model. The SVM approach tries to find the hyperplane that best discriminates the data. As stated before, our new vector is the composition of the projection of each window in the axis that best discriminates it. It also carries information about the reliability of each discriminant. It means that the SVM approach also matches the geometry of our data. The results of LDA+SVM method 1 are listed on Tables II-III.

LDA + SVM method 1 (350 to 1000 nm).						
Dataset	4 DPI (%)		7 DPI (%)		10 DPI (%)	
	TPR	SPC	TPR	SPC	TPR	SPC
Cohort 1 train	100.0	100.0	100.0	100.0	-	-
Cohort 1 *k-fold*	100.0	100.0	100.0	100.0	-	-
Cohort 2 head/thorace	100.0	100.0	100.0	100.0	63.6	100.0
Cohort 2 abdomen	100.0	100.0	96.1	100.0	42.8	100.0

TABLE II: LDA + SVM method 1 (350 to 1000 nm).

LDA + SVM method 1 F1 score (350 to 1000 nm).			
Dataset	4 DPI (%)	7 DPI (%)	10 DPI (%)
Cohort 2 head/thorace	100.0	100.0	77.75

TABLE III: LDA + SVM method 1 F1 score (350 to 1000 nm).

In the experiment using LDA+SVM method 2, we added one more feature to our model, generating $[x'_1, \ldots, x'_{26}, x'_{\text{all}}]$ as feature vector. The results of this approach are listed on Tables IV-V.

LDA + SVM method 2 (350 to 1000 nm).						
Dataset	4 DPI (%)		7 DPI (%)		10 DPI (%)	
	TPR	SPC	TPR	SPC	TPR	SPC
Cohort 1 train	100.0	100.0	100.0	100.0	-	-
Cohort 1 *k-fold*	100.0	89.5	100.0	92.9	-	-
Cohort 2 head/thorace	100.0	86.4	100.0	93.2	98.7	98.3
Cohort 2 abdomen	100.0	95.0	100.0	95.0	93.5	100.0

TABLE IV: LDA + SVM method 2 (350 to 1000 nm).

LDA + SVM method 2 F1 score (350 to 1000 nm).			
Dataset	4 DPI (%)	7 DPI (%)	10 DPI (%)
Cohort 2 head/thorace	92.70	96.48	98.49

TABLE V: LDA + SVM method 2 F1 score (350 to 1000 nm).

The addition of the LDA transformation of the entire interval from 350 to 1000 nm as a new feature to our model increased our generalization accuracy for the case of 10 days post infection, a case that we do not have in our training examples. This however comes with the cost of lowering our predictions for the cases of 4 and 7 days post infection.

Comparing both methods proposed in this work, Tables II and IV, method 2 generalizes better since the non-trained data of the mosquitoes with 10 DPI reached a higher performance. The reason is the insertion of the LDA representation of the whole spectrum as a new feature in the input vector. Although it increased the accuracy of the unseen scenario of 10 DPI, it generated a slight reduction in the performances for 4 and 7 DPI, SPC columns and lines Cohort 1 k-fold, Cohort 2 head/thorace, and Cohort 2 abdomen.

Observing the results presented in Tables I and IV, which are associated to the proposed LDA+SVM method 2 and the approach presented in [6], respectively, it is noticed that the proposed method achieved better balanced accuracy $\left(\frac{\text{TPR}+\text{SPC}}{2}\right)$ for all mosquitos stratifications, except for cohorts 1 k-fold and 2 head/thorace with 4 and 7 DPI's. It is worth noticing that the proposed technique performed better in non-trained 10 DPI data and similarly in the remaining data with a considerably reduced spectral range of (350, 1000] nm.

IV. Conclusion

The results presented in this work shows that it is possible to identify an infected mosquito with Zika virus using a reduced spectral information, which could possibly be implemented by a cheaper device than a professional spectrometer.

Comparing the RT-qPCR, the NIRS, and the proposed method technologies in this work, One can state that RT-qPCR is the technique currently used in arboviruses programs to detect mosquitoes with Zika virus. However, it has a high cost per measure, takes some time to obtain results, and requires specialized workers. Meanwhile, the NIRS approach reads wavelengths from 350 to 2500 nm, which requires a high-cost device, making it inviable to be used in the field. It has a low cost per measure, is fast to obtain results, and does not need specialized personnel. Finally, the proposed method reads wavelengths from 350 to 1000 nm, which requires a lower cost and reduced size device, making it is viable for usage in the field. It also has a low cost, is fast to obtain results, and does not need specialized workers.

The proposed LDA+SVM method 2 provided a better generalization in classifying mosquitoes within 10 DPI compared with the other methods analyzed in this work. The 10 DPI mosquitoes data is unseen since it was not applied in the training. The inclusion of the feature represented by the projection of the data in the LDA coordinates of the whole signals contributed to the generalization performance. The main point here is that the proposed method 2 reaches performance similar to the one in [6] with significantly reduced wavelength information.

Due to the difficulty of the task of capturing and infecting mosquitoes, our training dataset has a small number of samples, and to work with more reliable statistics information, it is recommended to expand the dataset, prefereably with different numbers of days post infection. This is supposed

to generate a better training model and possibly increase the system accuracy.

Further research on hyperparameters optimization and outliers removal is a natural future step to be followed, since the data is far from being exhaustively investigated. Future work also includes the construction a low budget IoT device to do these readings and to run our model on field.

REFERENCES

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*, volume 4. AMLBook, 2012.

[2] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.

[3] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In David Haussler, editor, *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 144–152, Pittsburgh, PA, USA, July 1992. ACM Press.

[4] S. A. Bustin. Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25(2):169–193, 2000.

[5] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, 2000.

[6] Jill N. Fernandes, Lílha M. B. dos Santos, Thaís Chouin-Carneiro, Márcio G. Pavan, Gabriela A. Garcia, Mariana R. David, John C. Beier, Floyd E. Dowell, Rafael Maciel de Freitas, and Maggy T. Sikulu-Lord. Rapid, noninvasive detection of zika virus in aedes aegypti mosquitoes by near-infrared spectroscopy. *Science Advances - American Association for the Advancement of Science (AAAS)*, 4(5), May 2018.

[7] Stan Z. Li and Anil Jain, editors. *LDA (Linear Discriminant Analysis)*, pages 899–899. Springer US, Boston, MA, 2009.

[8] A.M. Martinez and A.C. Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[9] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[10] C. Orlando, P. Pinzani, and M. Pazzagli. Developments in quantitative pcr. *Clinical chemistry and laboratory medicine*, 36(5):255–269, 1998.

[11] J. Papin, W. Vahrson, R. Hines-Boykin, and D. P. Dittmer. Real-time quantitative pcr analysis of viral transcription. *Methods in molecular biology*, 292(1):449–480, 2005.