ISSN 1870-4069

Classification of Zika and Dengue Clinical Data Using Feature Encoding and Machine Learning Techniques

Elí Cruz-Parada[1], Guillermina Vivar-Estudillo[2],
Eduardo Pérez-Campos[1], Carlos Lastre-Domínguez[1]

[1] Tecnológico Nacional de México, Instituto Tecnológico de Oaxaca,
División de Posgrado e Investigación,
Mexico

[2] Universidad Autónoma "Benito Juárez" de Oaxaca,
Facultad de Sistemas Biológicos e Innovación Tecnológica,
Mexico

Eli.cruz.parada@gmail.com,gvivar.cat@uabjo.mx,
{pcampos,carlos.lastre}@itoaxaca.edu.mx

Abstract. Dengue and Zika are viral diseases primarily transmitted to humans through bites from infected female Aedes aegypti mosquitoes, posing considerable medical concerns due to their potential severity. These illnesses can lead to fatal hemorrhagic fevers and have affected around 300 million people globally. Early diagnosis is crucial for effective treatment. Despite extensive research over recent decades, achieving diagnostic accuracy remains challenging. This study introduces a novel method for organizing clinical data to enhance the identification of Zika and Dengue by utilizing symptoms as features extracted from clinical studies and applying machine learning techniques for classification tasks. Rigorous statistical analysis using ANOVA and Kruskal-Wallis tests revealed p-values below 0.05, indicating significant findings. Additionally, the classifiers examined demonstrated AUCs and F1 scores exceeding 96%, highlighting their effectiveness. This approach aims to improve diagnostic precision, thereby facilitating timely intervention and reducing the impact of these diseases on global health.

Keywords: Tropical diseases, dengue, sika, clinical data, machine learning.

1 Introduction

Dengue and Zika fever are viral diseases transmitted to humans primarily through the bite of an infected female Aedes aegypti mosquito. While other species of the Aedes genus can also transmit these viruses, their role is generally secondary [1]. Once a mosquito becomes a vector, it remains so for its entire lifespan. Both diseases are more prevalent in tropical and subtropical regions, with an increasing incidence observed in recent years. The World Health Organization (WHO) reported a dramatic rise in dengue cases, from 505,430 cases in 2000 to 5.2 million in 2019 [2].

Dengue is often asymptomatic; however, when symptoms do appear, they typically resolve within one to two weeks. Despite being classified as mild or moderate, dengue

can develop into a severe form, involving bleeding and requiring hospitalization due to the risk of fatality. Common symptoms include fever, headache, retro-orbital pain, nausea, and vomiting [3]. A second dengue infection often leads to more severe illness, which can be misdiagnosed as other febrile illnesses. Similarly, Zika fever, caused by the ZIKV virus, is transmitted by the same mosquito species and can also be spread through sexual transmission.

Symptoms include rash, itching, non-purulent conjunctivitis, arthralgia, myalgia, and fever. Only about one in four infected individuals exhibit symptoms, which are generally mild and last for 2 to 7 days. The clinical presentation is often similar to dengue or Chikungunya, necessitating laboratory confirmation [4]. Currently, there is no specific treatment for dengue, with management focusing on pain relief, while avoiding nonsteroidal anti-inflammatory drugs (NSAIDs) due to bleeding risks.

Research teams worldwide are increasingly using machine learning and data mining techniques to improve disease diagnosis. For example, decision trees have been successfully used to differentiate between tropical infections [5]. In Paraguay, researchers achieved an average accuracy of 96% using Support Vector Machine classifiers and Artificial Neural Networks [6]. An Android application named GZC-DIAG outperformed resident physicians in diagnosing diseases with a 96.88% success rate [7].

Innovations continue with machine learning integrated into laboratory tests, such as peripheral blood smear (PBS) analysis, showing promising results with up to 95.74% accuracy in detecting Dengue Virus (DENV) [8]. Despite challenges related to data scarcity in specific clinical analyses, these advancements underscore the potential of AI in transforming diagnostic capabilities.

Moreover, beyond diagnosis, efforts are being made to predict the risks associated with diseases like dengue. Studies have demonstrated accuracies ranging from 70% to 96.27% using various techniques, including bioelectrical impedance analysis (BIA) and neural networks [9, 10]. These predictive models not only aid in diagnosing but also in assessing the prognosis and potential complications of patients.

Looking forward, the development of more affordable and portable diagnostic tools, such as biosensor devices, represents a significant advancement, particularly for resource-limited settings. Furthermore, AI-driven models have proven effective in forecasting disease outbreaks, achieving up to 89.25% accuracy in predicting dengue outbreaks [11]. In India, AI has been used to predict outbreaks and diagnose diseases like Zika using data from users and environmental factors. The processing time is 0.15 s with 91.25% accuracy [12].

The diagnosis by 3D super-resolution microscopy images has been used in Zika, these images are taken from the endoplasmic reticulum (ER). Deep learning techniques were able to identify morphological changes in the ER caused by the virus [13]. Similarly, ensemble methods were applied to identify cases of congenital Zika, these were based on the U.S. Zika Pregnancy and Infant Registry (USZPIR) and the Zika Active Pregnancy Surveillance System (ZAPSS) of Puerto Rico, and although it presented a high sensitivity (96% for USZPIR and 97% for ZAPSS), the model was specifically designed for this dataset [14].

Research on using Electrocardiogram (ECG)-derived heart rate variability (HRV) metrics and machine learning (ML) models to predict infant exposure to Zika virus (ZIKV) has been conducted. In a study of 21 infants with an average age of 15 months, a cubic support vector machine classifier was applied to their ECGs [15]. The research

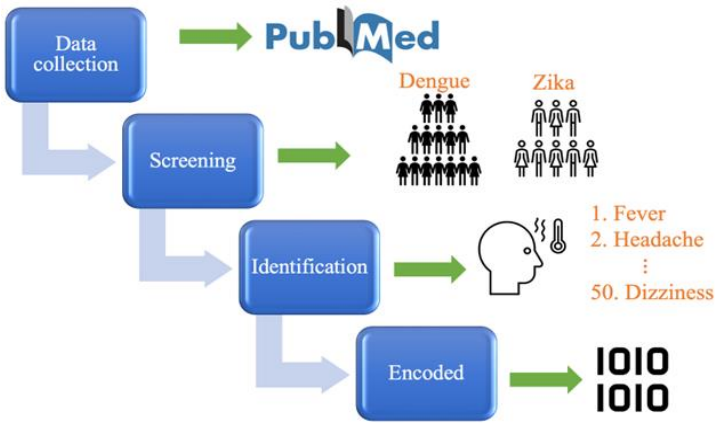

Fig.1. Processing system for a codified symptoms database.

Table 1. Codification of symptoms.

# Patient	Pathology	Fever	Headache	Myalgia	Nausea	Rash
1	Zika	0	0	0	0	1
2	Dengue	1	1	0	0	0
3	Dengue	1	1	1	0	0
4	Zika	0	1	0	0	1

team reported that their model was able to differentiate between infants affected by Zika, non-affected by Zika, and those not affected by the virus with a predictive value of 86%. However, there is some dispute about whether HRV is a specific attribute of Zika.

A review, reported in 2022, concluded that machine learning and deep learning techniques for diagnosing arboviral diseases focus mainly on dengue and do not effectively differentiate between more than two different pathologies. It was also noted that decision tree-based techniques are the most used [16].

The approaches mentioned above yield important findings, but they are not sufficient for accurately diagnosing tropical diseases. We propose a feature coding method that takes into account the 50 most common symptoms and utilizes advanced machine-learning techniques to identify Zika and Dengue pathology. This work comprises several sections. Section 2 provides a brief overview of the database. Section 3 offers an in-depth analysis of features using ANOVA and Kruskal Wallis, including p-value analysis. Section 4 outlines the performance of classifiers in determining Zika and Dengue. Section 5 discusses the results of this work. Finally, in Section 6, we present our conclusions.

2 Database

As depicted in Fig. 1, we have developed a methodology for establishing a database consisting of four subprocesses: data collection, screening, identification, and encoding. We will now elaborate on each stage.

2.1 Data Collection and Screening

To gather reliable data and establish a database, a search was conducted in the PubMed database using the keywords "dengue," "prevalence," and "clinical symptoms." The selected studies had to meet specific criteria. Firstly, they needed to report on at least 50 patients diagnosed with dengue through laboratory tests. Secondly, they had to mention at least five symptoms observed in confirmed dengue patients. Thirdly, they were required to provide the number of patients affected by each symptom, including the mortality rate. Finally, they needed to confirm that the symptoms were observed within the first four days of the disease.

To gather Zika data, a search was conducted in the PubMed database using the keywords "Zika fever," "prevalence," and "clinical symptoms." However, as there were significantly more papers for Dengue (1,999 papers) than for Zika (673 papers), the selection criteria were adjusted as follows:

1. The study must report on at least 20 patients diagnosed with Zika through laboratory tests.
2. The study should mention more than five different symptoms observed in confirmed Zika patients.
3. The study should provide the number of patients who experienced each symptom.
4. The study should present the mortality rate.
5. The symptoms should have been observed within the first four days of the disease.

The following criteria are used to select papers: Each paper is from relevant journals, such as *PLOS Neglected Tropical Diseases* [17], *BMC Infectious Diseases* [18], *Annals of Medicine* [19], *The Lancet Infectious Diseases* [20], among others. We aim to minimize redundant investigations to ensure that each study provides unique and valuable information. To avoid bias, data is selected from different hospitals, years, and countries for each study.

2.2 Identification

For the identification processing, the age, gender, economic status, and nationality of the patients were not recorded in any of the cases. After collecting the necessary data, 22,379 dengue-confirmed patients and 7,135 Zika-confirmed patients were observed for up to 37 and 34 different symptoms, respectively. In total, 20 common and 30 other symptoms were observed, making up 50 symptoms.

2.3 Encoded

The patients' symptoms were coded, a label was added to each symptom indicating “1” if the symptom was present and “0” if it was absent. To illustrate the codification, table 1 considers four patients and five symptoms: Patient 1 is coded as 00001 with Zika pathology. Subsequently, we established a database coded for Zika and Dengue symptoms. Next, a statistical analysis is performed to find significance.

Table 2. F-value and H-value calculated for some symptoms.

Symptoms	ANOVA	Kruskal-Wallis
Rash	1.6871×10^4	1.0735×10^4
Fever	1.4924×10^4	9.9118×10^3
Conjunctivitis	6.4686×10^3	5.3059×10^3
Pruritus	3.1877×10^3	2.8770×10^3
Low back pain	2.3153×10^3	2.1469×10^3
Retro orbital pain	1.6269×10^3	1.5419×10^3
Sore throat	1.1335×10^3	1.0916×10^3
Edema	1.1249×10^3	1.0836×10^3

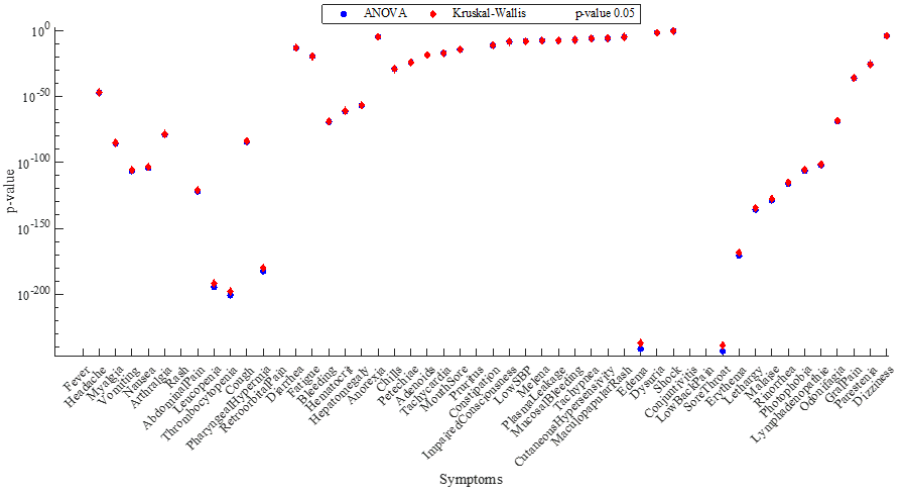

Fig.2. p-value with ANOVA and Kruskal-Wallis analysis.

3 Features Statistical Analysis

3.1 ANOVA

The features are evaluated through the ANOVA test, calculating the F-value by mean equation (1):

$$F = \frac{MSB}{MSW}. \tag{1}$$

To calculate the mean square between (MSB) groups for the ANOVA test, it should start by calculating the mean of each group.

Then, calculate the overall mean (the mean of all data points combined), and finally calculate the sum of squares between groups (SSB) by equation (2).

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \tag{2}$$

where n_i is the number of observations in group i, $\bar{X}_i$ is the mean of group i and $\bar{X}$ is the overall mean. The degrees of freedom between groups (df_b) can be computed by equation (3):

$$df = k - 1, \tag{3}$$

where k is the number of groups. Calculated the SSB and df_b, the MSB is computed by equation (4):

$$MSB = \frac{SSB}{df_b}. \tag{4}$$

To compute mean square within (MSW) groups for the ANOVA test, initially, the sum of squares within groups (SSW) is calculated by equation (5):

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \tag{5}$$

where X_{ij} is the j-th observation in group i. The degrees of freedom within (df_w) are calculated by equation (6):

$$df_w = N - 1, \tag{6}$$

where N is the total number of observations across all groups. Finally, the MSW is calculated by the equation (7):

$$MSW = \frac{SSW}{df_w}. \tag{7}$$

3.2 Kruskal-Wallis Test

A Kruskal-Wallis analysis was also conducted using the statistical test (H or χ^2 chi-square) represented by equation (11):

$$H = \frac{12}{N_s\,(N_s + 1)} \sum_{i=1}^{k} \left(\frac{SS_i}{n_i}\right) - 3(N_s + 1), \tag{8}$$

where k is the number of groups, N_s is the total number of observations, n_i is the number of observations in the i-th group and SS_i is the sum of the squared ranks within the i-th group.

In Table 2, ANOVA and Kruskal-Wallis analyses are being performed to compare symptoms such as rash, fever, conjunctivitis, pruritus, low back pain, retro-orbital pain, sore throat, and edema. As can see in Fig. 2, the statistical analysis provides a detailed p-value analysis to determine the statistical significance of these symptoms, with the majority showing a p-value of less than 0.05.

4 Classification

We are evaluating five classifiers: cubic SVM, quadratic SVM, Gaussian, fine KNN, and weighted KNN. We conducted an AUC analysis to assess the overall discriminative ability of the classifiers between the positive and negative classes. The AUC is determined by the true positive rate (TPR) and the false positive rate (FPR), which are calculated by equation (9) and equation (10), respectively:

$$TPR = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + Falses\ Positives\ (FP)}, \tag{9}$$

$$FPR = \frac{False\ Positives\ (FP)}{False\ Positives\ (FP) + True\ Negatives\ (TN)}. \tag{10}$$

As seen in Fig. 3, the classifiers achieve an AUC near 0.96, indicating a high level of performance.

The F1-score was calculated, and the accuracy of the test was measured using equation (11):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{11}$$

where precision and recall are calculated by equation (12) and (13), respectively:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FN)}, \tag{12}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}. \tag{13}$$

Under 100 iterations and a holdout of 80/20, we performed calculations to determine the average Area Under the Curve (AvgAUC) and its standard deviation (StdAUC), as well as the average F1-Score (AvgF1) and its standard deviation (StdF1) for balanced sets, with the minority class being the Zika pathology. Subsampling was conducted for the majority class, resulting in three sets.

The first and second sets consisted of 7135 data for Zika and Dengue, respectively, while the third set comprised 8109 data for Dengue and 7135 for Zika, with the majority class being dengue. Tables 3, 4, and 5 display the performance of different classifiers, all of which exceeded the 96% threshold.

5 Discussion

The effectiveness of machine learning techniques also depends on how the data is represented. Therefore, it is essential to conduct a statistical analysis to distinguish between Zika and Dengue based on their symptoms. After performing ANOVA and Kruskal-Wallis tests, it was determined that symptoms such as fever, rash, conjunctivitis, pruritus, low back pain, retro-orbital pain, sore throat, and edema exhibited the most statistical significance.

Table 3. AUC Average performance for the first set.

Classifier	AvgAUC	StdAUC	AvgF1	StdF1
Fine KNN	0.9689	0.00285	0.9683	0.0029503
Cubic SVM	0.9862	0.00282	0.9817	0.0019973
Medium Gaussian SVM	0.9870	0.00318	0.9824	0.0018754
Quadratic SVM	0.9871	0.00344	0.9821	0.0018718
Weighted SVM	0.9758	0.00266	0.9626	0.0030827

Table 4. AUC Average performance for the second set.

Classifier	AvgAUC	StdAUC	AvgF1	StdF1
Fine KNN	0.9703	0.00260	0.9694	0.00276
Cubic SVM	0.9865	0.00299	0.9816	0.00228
Medium Gaussian SVM	0.9890	0.00269	0.9822	0.00219
Quadratic SVM	0.9750	0.00302	0.9823	0.00198
Weighted SVM	0.9758	0.00266	0.9618	0.00303

Table 5. AUC Average performance for the third set.

Classifier	AvgAUC	StdAUC	AvgF1	StdF1
Fine KNN	0.96869	0.00330	0.9676	0.00351
Cubic SVM	0.98496	0.00296	0.9813	0.00265
Medium Gaussian SVM	0.98591	0.00290	0.9818	0.00250
Quadratic SVM	0.98801	0.00355	0.9816	0.00248
Weighted SM	0.97419	0.00287	0.9609	0.00378

Therefore, these symptoms are crucial for predicting diseases using machine learning techniques. Despite utilizing 50 symptoms, the computational cost was not significantly affected.

While this did not significantly impact the current research, adding data from additional diseases could substantially increase computational costs.

The study also identified the best-performing classifiers. Decision tree-based classifiers and ensemble models like random forest, adaboost, and gradient boosting exhibited acceptable performance levels of around 90%, although they were found to be less effective than SVM-based models and certain non-parametric algorithms such as k-nearest neighbors (KNN). This study specifically focused on a limited number of diseases, namely Zika and Dengue. If additional diseases are included, it will be necessary to analyze more symptoms and conduct more rigorous statistical analyses.

6 Conclusions

We have developed a new method for organizing clinical data to help detect diseases such as Zika and Dengue. Our analysis revealed significant differences, with a p-value of 0.05, using ANOVA and Kruskal-Wallis tests. When we tested the features using various classifiers, we achieved an average performance of over 96%, ranging from 96% to 99%, through iterative training and testing. The classification was performed using an 80/20 holdout and 100 iterations, and the results from the classifiers demonstrate that the data is well represented.

In our future research, we aim to analyze more clinical data and explore other tropical diseases like Chikungunya. We also plan to use different machine learning models for classification and investigate the incorporation of deep learning techniques.

References

1. Malavige, G. N., Fernando, S., Fernando, D. J., Seneviratne, S. L. : Dengue viral infections. Postgraduate Medical Journal, vol. 80, no. 948, pp. 588–601 (2004) doi: 10.1136/pgmj.2004.019638
2. World Health Organization: Dengue and severe dengue. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, last accessed 2023/05/10
3. Guo, C., Zhou, Z., Wen, Z., Liu, Y., Zeng, C., Xiao, D., Ou, M., Han, Y., Huang, S., Liu, D., Ye, X., Zou, X., Wu, J., Wang, H., Zeng, E. Y., Jing, C., Yang, G.: Global epidemiology of dengue outbreaks in 1990-2015: A systematic review and meta-analysis. Frontiers in cellular and infection microbiology, vol. 7, no. 317 (2017) doi: 10.3389/fcimb.2017.00317
4. World Health Organization: Zika virus. https://www.who.int/news-room/fact-sheets/detail/zika-virus, last accessed 2023/05/10
5. Shenoy, S., Rajan, A. K., Rashid, M., Chandran, V. P., Poojari, P. G., Kunhikatta, V., Acharya, D., Nair, S., Varma, M., Thunga, G.: Artificial intelligence in differentiating tropical infections: A step ahead. PLoS neglected tropical diseases, vol. 16, no. 6 (2022)
6. Mello-Román, J. D., Mello-Román, J. C., Gómez-Guerrero, S., García-Torres, M.: Predictive models for the medical diagnosis of dengue: A case study in Paraguay. Computational and Mathematical Methods in Medicine (2019) doi: 10.1155/2019/7307803
7. Rodríguez-De-Araújo, A. P., Macedo-de-Araujo, M. C., Coutinho-Cavalcanti, T., de-Lacerda-Vidal, C. F., Gomes-Netto-Monte-da-Silva, M.: DZC DIAG: mobile application based on expert system to aid in the diagnosis of dengue, zika, and chikungunya. Medical & Biological Engineering & Computing, vol. 58, no. 11, pp. 2657–2672 (2020) doi: 10.1007/s11517-020-02233-6
8. Mayrose, H., Bairy, G. M., Sampathila, N., Belurkar, S., Saravu, K.: Machine learning-based detection of dengue from blood smear images utilizing platelet and lymphocyte characteristics. Diagnostics (Basel, Switzerland), vol. 13, no. 2, pp. 220 (2023) doi: 10.3390/diagnostics13020220
9. Faisal, T., Ibrahim, F., Taib, M. N.: A noninvasive intelligent approach for predicting the risk in dengue patients. Expert Systems with Applications, vol. 37, no. 3, pp. 2175–2181 (2010) doi: 10.1016/j.eswa.2009.07.060
10. Ibrahim, F., Faisal, T., Salim, M. I., Taib, M. N.: Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. Medical & biological engineering & computing, vol. 48, no. 11, pp. 1141–1148 (2010) doi: 10.1007/s11517-010-0669-z

11. Anggraini-Ningrum, D. N., Yu-Chuan, J. L., Hsu, C. Y., Solihuddin-Muhtar, M., Pandu-Suhito, H. P.: Artificial intelligence approach for severe dengue early warning system. Stud Health Technol Inform, pp. 881–885 (2023) doi:10.3233/SHTI231091
12. Dadheech, P., Mehbodniya, A., Tiwari, S., Kumar, S., Singh, P., Gupta, S., Atiglah, H. K.: Zika virus prediction using AI-driven technology and hybrid optimization algorithm in healthcare. Journal of Healthcare Engineering, vol. 12, pp. 2793850 (2022) doi: 10.1155/ 2022/2793850
13. Long, R. K., Moriarty, K. P., Cardoen, B., Gao, G., Vogl, A. W., Jean, F., Nabi, I. R.: Super resolution microscopy and deep learning identify zika virus reorganization of the endoplasmic reticulum. Scientific Reports, vol. 10, no. 1, p. 20937 (2020) doi: 10.1038/ s41598-020-77170-3
14. Lusk, R., Zimmerman, J., van-Maldeghem, K., Kim, S., Roth, N. M., Lavinder, J., Fulton, A., Raycraft, M., Ellington, S. R., Galang, R. R.: Exploratory analysis of machine learning approaches for surveillance of zika-associated birth defects. Birth Defects Research, vol. 112, no. 18, pp. 1450–1460 (2020) doi: 10.1002/bdr2.1767
15. Herry, C. L., Soares, H. M. F., Schuler-Faccini, L., Frasch, M. G.: Machine learning model on heart rate variability metrics identifies asymptomatic toddlers exposed to zika virus during pregnancy. Physiological Measurement, vol. 42, no. 5 (2021) doi: 10.1088/1361-6579/ac010e
16. da Silva-Neto, S. R., Tabosa-Oliveira, T., Teixeira, I. V., Aguiar de Oliveira, S. B., Souza-Sampaio, V., Lynn, T., Endo, P. T.: Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review. PLoS Neglected Tropical Diseases, vol. 16, no. 1 (2022) doi: 10.1371/journal.pntd.0010061
17. Rafi, A., Nahar-Mousumi, A., Ahmed, R., Haque-Chowdhury, R., Wadood, A., Hossain, G.: Dengue epidemic in a nonendemic zone of Bangladesh: Clinical and laboratory profiles of patients. PLOS Neglected Tropical Diseases, vol. 14, no. 10 (2020) doi: 10.1371/journal. pntd.0008567
18. Hasan, M. J., Tabassum, T., Sharif, M., Khan, M. A. S., Bipasha, A. R., Basher, A., Amin, M. R.: Comparison of clinical manifestation of dengue fever in Bangladesh: an observation over a decade. BMC infectious diseases, vol. 21, no. 1, pp. 1113 (2021) doi; 10.1186/s1287 9-021-06788-z
19. Asaga-Mac, P., Tadele, M., Airiohuodion, P. E., Nisansala, T., Zubair, S., Aigohbahi, J., Panning, M.: Dengue and zika seropositivity, burden, endemicity, and cocirculation antibodies in Nigeria. Annals of Medicine, vol. 55, no. 1, pp. 652–662 (2023) doi: 10.1080/ 07853890.2023.2175903
20. Marc-Ho, Z. J., Chanditha-Hapuarachchi, H., Barkham, T., Chow, A., Ng, L. C., Vernon-Lee, J. M., Sin-Leo, Y., Prem, K., Georgina-Lim, Y. H., de-Sessions, P. F., Rabaa, M. A., Seng-Chong, C., Hua-Tan, C., Rajarethinam, J., Tan, J., Anderson, D. E., Ong, X., Cook, A. R., Chong, C. Y., Hsu, L. Y., et. al.: Outbreak of zika virus infection in Singapore: an epidemiological, entomological, virological, and clinical analysis. The Lancet Infectious Diseases, vol. 17, no. 8, pp. 813–821 (2017)