Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 13 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i13.50437

PAPER

Predictive Analysis of Vector-Borne Diseases through Tabular Classification of Epidemiological Data

Orlando Iparraguirre-Villanueva[1](✉), Michael Cabanillas-Carbonell[2]

[1]Facultad de Ingeniería, Universidad Tecnológica del Perú, Chimbote, Perú

[2]Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú

oiparraguirre@ieee.org

ABSTRACT

Vector-borne diseases (VBDs) are major threats to human health. They are estimated to cause more than 700,000 deaths each year. This presents serious health problems for CBD. In recent years, the incidence of VBDs has increased globally, affecting one billion people approximately and accounting for 17% of all infectious diseases. Globally, disease rates have risen at an alarming rate, with more than 3.9 billion people at risk of infection. Therefore, it is essential to find approaches to detect these diseases; this is where machine learning (ML) models come into play. The purpose of this study was to predict VBDs using tabular epidemiological data. For this purpose, a set of ML models was used, such as support vector classifier (SVC), extreme gradient boosting (XGBoost), LightGBM, CatBoost, random forest (RF), and balanced random forest (BRF). A dataset consisting of 65 features and 1262 records was used during the training stage. The results highlighted the successful integration of the different models, such as SVC, XGBoost, LightGBM, CatBoost, BRF, and RF, with weights of 0.49959 ± 0.27112, 0.58496 ± 0.22619, 0.48482 ± 0.29971, 0.54992 ± 0.27982, 0.24924 ± 0.22654, and 0.45592 ± 0.25849. In addition, the BRF model stood out for having the lowest log loss, evaluated through the ensemble log-loss metric, with an average of 0.24924 and a standard deviation of 0.22654.

KEYWORDS

prediction, machine learning (ML), evaluation, models

1 INTRODUCTION

Vector-borne diseases (VBDs), such as malaria, dengue fever, and West Nile virus (WNV), pose significant threats to human health [1]. These diseases are spread to the population by infected insects or direct human-to-human transmission [2]. It is estimated that they cause more than 700,000 deaths a year globally [3]. According to the World Health Organization (WHO), 250,000 deaths a year due to malnutrition, heat stress, and VBDs are projected to occur in the coming years [4]. This highlights the significant health problems that VBDs represent [5]. In recent years, the incidence of VBDs has increased worldwide [6], affecting approximately one billion people

Iparraguirre-Villanueva, O., Cabanillas-Carbonell, M. (2024). Predictive Analysis of Vector-Borne Diseases through Tabular Classification of Epidemiological Data. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(13), pp. 103–117. https://doi.org/10.3991/ijoe.v20i13.50437

Article submitted 2024-06-05. Revision uploaded 2024-07-11. Final acceptance 2024-07-14.

and accounting for 17% of all infectious diseases [7]. These diseases are particularly high in tropical and subtropical regions, disproportionately affecting poor communities without access to drinking water [8]. Globally, rates have shown worrying growth, putting more than 3.9 billion people at risk of infection [9]. These diseases present significant challenges to human health [10]. Dengue, malaria, and chikungunya are examples of extremely harmful VBDs, resulting in millions of deaths annually [11]. According to the WHO, 93% of the population in India is at risk of contracting malaria [12]. Tick-borne diseases are expanding geographically and in incidence, especially in temperate regions of Europe, where ticks are zoonotic vectors of public health importance [13]. In the United States, more than 30,000 new cases of Lyme disease are reported annually, although this number is estimated to represent only 10% of total cases due to underreporting and underdiagnosis [14]. The increase in VBDs threatening humans is evident [15]. Most VBDs are not vaccine-preventable and can only be controlled with integrated interventions, including vector control [16]. Managing these diseases would contribute to improving the health and well-being of human beings [17].

Over the past few decades, the rapid development of artificial intelligence (AI) technologies, such as machine learning (ML) and deep learning (DL), has led to significant transformation in various industries, with a particular impact on medical diagnoses [18]. These innovations, applied to intelligent diagnosis, image interpretation, and accurate disease classification, have revolutionized the field of medicine [19]. AI has been integrated into various medical disciplines to assist professionals in the diagnostic process [20]. AI is also expected to bring significant societal benefits, including the potential to eradicate disease [21]. These technological advances are also being used to advance the field of medicine [22], processing patient data automatically, and improving the accuracy of predictions [23]. In the current era of big data, AI is positioned as an important resource that offers new opportunities to address health problems and optimize patient care [24]. AI is becoming an important tool for addressing urgent health problems [25]. AI provides new ways to analyze and predict patient data using algorithms that automate information processing [26]. This capability represents a remarkable advance in the accuracy and efficiency of medical practice.

The study aimed to create a system to predict VBDs using tabular epidemiological data. For this purpose, a set of ML models (SVC, XGBoost, LightGBM, CatBoost, RF, and BRF) were used. The first two parts of this study are composed of the background and a review of related literature. The third part details the methodology employed in the study. The model training results are presented in the next section. Finally, the last sections address the discussions and conclusions of the study.

2 RELATED WORK

Study-related follows. For example, the authors of [27] used artificial neural networks (ANNs) and support vector machines (SVMs) to predict dengue fever, achieving an accuracy of 96.19%. Similarly, in [28], an ML approach was developed to predict dengue fever employing decision tree (DT) and RF algorithms. DT achieved an outstanding accuracy of 79%. In [29], they compared various ML models such as logistic regression (LR), SVM, DT, and RF to compare their accuracy in predicting dengue prevalence. The SVM model achieved the best performance with an accuracy of 76%.

On the other hand, the authors of [30] developed two malaria classifiers to predict the presence of the disease in patients with and without history. They used ML techniques such as neural networks (NN), LR, SVM, and k-nearest neighbors (KNN). The LR classifier showed superior performance with an accuracy of 97.14%. In [31], they applied various ML techniques (LR, DT, SVM, and RF) for malaria prediction. The LR model exhibited the best performance, achieving an accuracy of 83% and an F1 score of 90%. In [32], ML models such as XGBoost, KNN, SVM, DT, LR, RF, naïve Bayes (NB), AdaBoost, and explainable boosting machines (EBM) were implemented. The RF and EBM models outperformed other models with an accuracy of 84%. In [33], various ML techniques such as RF, SVM, and ANN were explored for malaria prediction. RF obtained the highest accuracy with 92%.

Likewise, in research [34], spatial prediction of cutaneous leishmaniasis was performed using three ML algorithms: DT, SVR, and LR. The results were satisfactory, with 0.951, 0.934, and 0.914 accuracy values for the DT, SVR, and LR algorithms.

3 METHODOLOGY

First, a detailed description of the models (SVC, XGBoost, LightGBM, CatBoost, BRF, and RF) used to predict VBDs is presented. A comprehensive dataset analysis is carried out in the second and final part.

3.1 Description of ML models

Support vector classification: This classification technique uses two parallel hyperplanes to distinguish between two classes of data, maximizing the margin between them [35]. This method, derived from the principle of structural risk minimization, is based on statistical learning theory and is widely used for pattern classification [36]. It is considered an efficient and simple classification method [37]. The model can be expressed in equations (1) and (2).

$$\min 1/2w^2, \tag{1}$$

Which is subject to:

$$y_i(wx + b) - 1 \geq 0, i = 1 \ldots .n, \tag{2}$$

Where b is the bias term, w is the weight vector, x is the vector characteristic, n is the number of samples, and y_i is the class label of the sample.

Extreme gradient boosting: XGBoost is an ML algorithm that stands out in integrating multiple regression trees using the boost method [38]. The XGBoost technique involves using boost to learn multiple decision trees iteratively. It starts by training one tree to predict the outcome, and then the next tree is trained based on the residuals obtained. XGBoost makes an early prediction by assigning a value to the tree root. The residuals are calculated using this first predicted value, i.e., the difference between the predicted and observed values in the dataset. All of these residuals are then mapped to the tree root [39]. Equation (3) expresses the mathematical equation of the model.

$$\hat{y}_i = \sum_{t=1}^{m} f_t(x_i), \tag{3}$$

In this formula, the variable *y* represents the anticipated overall projection of the model, while *f(x)* denotes the value estimated by the decision tree at position *i*.

LightGBM: It is an algorithm designed to be lightweight and efficient, standing out for maintaining high accuracy [40]. It is a histogram-based decision tree algorithm and is presented as an innovative ML technique centered on joint trees, based on the gradient-boosted decision tree approach [41]. The model can be expressed in equation (4).

$$y = \sum_{i=1}^{N} f_i(x), \tag{4}$$

In the given context, $f_i(x)$ represents the prediction by the *i*-th decision tree based on the characteristics *x*, whereas *y* is the overall prediction of the model, and *N* denotes the total number of trees present in the model.

CatBoost: The core idea of the CatBoost algorithm lies in the ability to generate a robust regression model by iteratively combining weak regressors [42]. CatBoost is a Gradient Boosting Decision Trees algorithm based on balanced tree construction and was introduced in 2018 to cope with prominent challenges in machine learning models. These challenges include the inability to handle categorical variables directly and the susceptibility of models to overfitting [43]. CatBoost has several unique features, most notably its ability to directly handle categorical characteristics without needing one-hot coding or other types of feature preprocessing. Equation (5) describes the model.

$$\hat{y} = \sum_{\kappa=1}^{K} f_\kappa(x), \tag{5}$$

Where, $\hat{y}$ is the model prediction, K is the total number of trees in the set, and $f_K(x)$ is the prediction of the *k*-th tree.

Balanced random forest: This technique stands out for its ability to improve the accuracy of minority class identification. This variant, derived from RF, follows a strategy where, for each tree, two bootstrapped sets of the same size are created, equivalent to the size of the minority class. One of these sets is intended for the minority class, and the other for the majority class, together forming the training set [44]. The main goal is to maximize predictive power, ensuring that all training data is used in the construction of the classification model [45].

Random forest: RF is characterized by being composed of numerous trees [46]. This algorithm operates by taking several samples from the original data, constructing numerous uncorrelated decision trees with different training samples, and calculating the mean or classifying all decision trees. The result of this operation is used to perform regression or classification [47]. In the ML domain, RF Random stands out as a powerful algorithm [48]. The model uses the formula in Equation (6).

$$\hat{y} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} f_i(x) \tag{6}$$

Where N is the total number of trees and $f_i(x)$ is the prediction of the *i*-th tree.

In the ML domain, the performance evaluation of classification models is critical to determining their effectiveness.

Therefore, we employ several metrics for this work that provide relevant information about their capability. Accuracy, recall, specificity, and F1 score are the most used metrics. A detailed explanation of each is presented below.

Accuracy is the proportion of correct positive predictions out of all the predictions made by the model.

Recall, also known as sensitivity, measures the ability of the model to identify all true positive instances.

Specificity focuses on the model's ability to classify negative cases correctly.

F1-Score is used as a composite measure to balance accuracy and count.

MAP@3 (mean average precision at 3) is a ranking metric that assesses the model's capability to rank instances properly according to their importance. A high MAP@3 value signifies the effectiveness of the model.

Logloss (logarithmic loss) is a loss function used to evaluate the performance of probabilistic classification models.

3.2 Case study

Understanding the dataset: This study uses a dataset from the Kaggle repository comprising 1262 records and 65 variables. The columns of the dataset represent various health-related variables. For example, the column 'sudden_fever' measures the presence or absence of sudden fever, while 'headache' refers to this symptom, 'mouth_bleed' refers whether or not there is bleeding in the mouth, 'nose_bleed' indicates whether or not there is a nosebleed, 'muscle_pain' refers to the presence or absence of muscle pain, 'joint_pain' indicates whether or not there is joint pain, 'vomiting' represents the presence or absence of vomits, 'rash' indicates the presence or absence of skin rashes, 'diarrhea' refers to the presence or absence of diarrhea, 'hypotension' indicates whether or not there is hypotension (low blood pressure), 'pleural_effusion' indicates the presence of fluid accumulation in the pleural cavity, 'ascites' refers to abnormal fluid accumulation in the abdominal cavity, 'gastro_bleeding' indicates the presence of gastrointestinal bleeding, 'swelling' represents whether or not there is swelling, 'nausea' indicates the presence of it, 'chills' refers to the sensation of shivering, 'myalgia' indicates the presence of muscle pain, 'digestion_trouble' refers to problems in digestion, 'fatigue' indicates the presence of it, 'skin_lesions' represents whether or not there are skin lesions, 'stomach_pain' indicates its presence or absence, 'orbital_pain' refers to the presence or absence of pain in the eye socket, 'neck_pain' indicates whether or not there is neck pain, 'weakness' represents the presence or absence of weakness, 'back_pain' indicates whether or not there is back pain, 'weight_loss' refers to whether or not there is weight loss, 'gum_bleed' indicates whether or not there is bleeding gums, 'jaundice' represents jaundice or yellowing of the skin, 'coma' indicates whether or not there is coma, 'diziness' refers to the feeling of dizziness, 'inflammation' indicates the presence of it, 'red_eyes' represents whether or not there is redness of the eyes, 'loss_of_appetite' indicates loss of appetite, 'urination_loss' refers to loss of urine, 'slow_heart_rate' indicates the presence of a slow heart rate, 'abdominal_pain' refers to the presence or absence of abdominal pain, 'light_sensitivity' indicates whether there is sensitivity to light, 'yellow_skin' indicates whether there is yellowing of the skin, 'yellow_eyes' indicates the presence of yellow eyes, 'facial_distortion' refers to whether or not there is facial distortion, 'microcephaly' indicates whether or not there is microcephaly, 'rigor' represents muscle stiffness, 'bitter_tongue' indicates whether or not there is

bitter taste on the tongue, 'convulsion' indicates the presence of convulsions, 'anaemia' indicates the presence of anaemia, 'cocacola_urine', 'hypoglycaemia' indicates low blood sugar, 'prostration' indicates prostration, 'hyperpyrexia' indicates high fever, 'stiff_neck' indicates neck stiffness, 'irritability' indicates this symptom, 'confusion' indicates the existence of mental confusion, 'tremor' indicates the existence of tremors, 'paralysis' indicates the existence of paralysis, 'lymph_swells' indicates whether or not there is swelling of the lymph nodes, 'breathing_restriction' indicates the presence of breathing restriction, 'toe_inflammation' indicates the presence of swelling of the toes, 'finger_inflammation' refers to the presence of swelling of the fingers, 'lips_irritation' indicates whether or not there is irritation of the lips, 'itchiness' indicates the presence of itching sensation, 'ulcers' indicates the presence of ulcers, 'toenail_loss' indicates whether or not there is loss of toenails, 'speech_problem' indicates whether or not there are speech problems, 'bullseye_rash' indicates the presence of skin rash, and finally, the variable 'prognosis' which is related to the medical diagnosis. The development of the study case is presented in Figure 1.

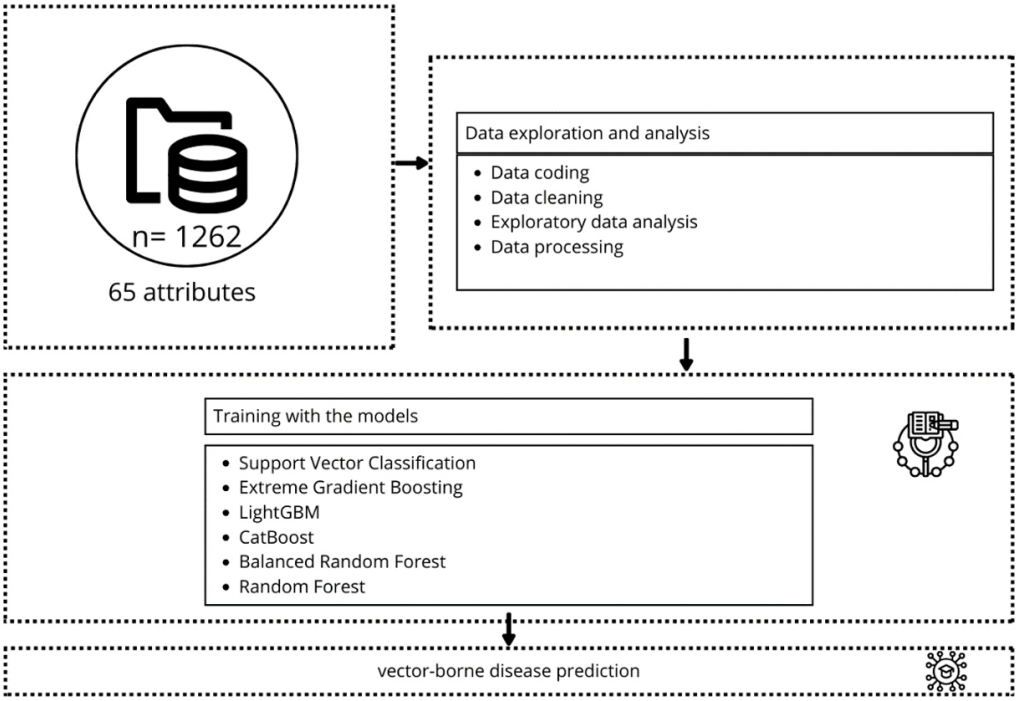

Fig. 1. Development process of the case study

Data preparation: The dataset was evaluated before the training of the models. First, the NumPy, pandas, and matplotlib libraries were imported for dataset manipulation and visualization. Seaborn was also used to adjust the color palette. Then, 'is_generated' columns were added to the training and test sets with assigned values. An index adjustment was performed for the original data, and a new 'id' column was created. The 'prognosis' column was modified in the original data, replacing spaces with underscores. Subsequently, the training sets and original data were concatenated into 'df_concat'. The target column is specified as 'prognosis'. Finally, the first rows of the test ('df_test') and original data ('original') sets are presented for inspection and verification. These actions are essential to ensure data consistency, completeness, and relevance (refer to Table 1).

Table 1. Variables of the dataset

#	Sudden_Fever	Headache	Mouth_Bleed	Nose_Bleed	...	Prognosis	Is_Generated
0	1	1	0	1	...	Lyme_disease	1
1	0	0	0	0	...	Tungiasis	1
2	0	1	1	1	...	Lyme_disease	1
3	0	0	1	1	...	Zika	1
4	0	0	0	0	...	Rift_Valley_fever	1

Exploratory analysis of the data: Figure 2 illustrates the frequency of various symptoms associated with vector-borne diseases. According to Figure 2a, fever or headache is observed to be the most prevalent symptom, registering more cases. Figure 2b shows that mouth bleeding is a frequent symptom among patients. Also, Figure 2c shows that most patients present muscle pain. Similarly, Figure 2d shows that vomiting is a frequent symptom among patients. This visual analysis provides a quantitative representation of the distribution of symptoms in the population.

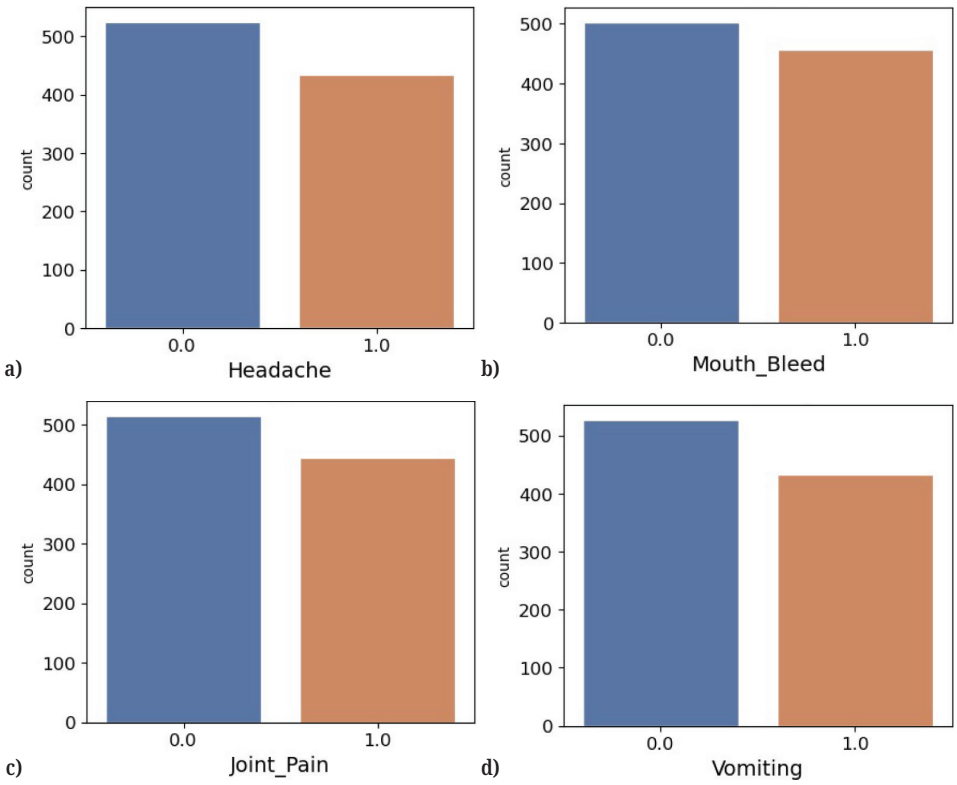

Fig. 2. The main symptoms: (a) fever or headache is the most prevalent symptom, (b) mouth bleeding, a frequent symptom, (c) muscle pain, and (d) vomiting is a common symptom in patients

Below is a graph representing the percentage distribution of the various VBDs in the dataset. According to Figure 3, 7% of the cases are Lyme disease, 8% are yellow fever, 8% are dengue fever, and 8% are plague. In addition, 9% of cases are for chikungunya, 9% for malaria, 9% for West Nile fever, 10% for Rift Valley fever, 10%

for Zika, and 11% for tungiasis. These percentages offer a visual depiction of the distribution of each disease in the analyzed dataset.

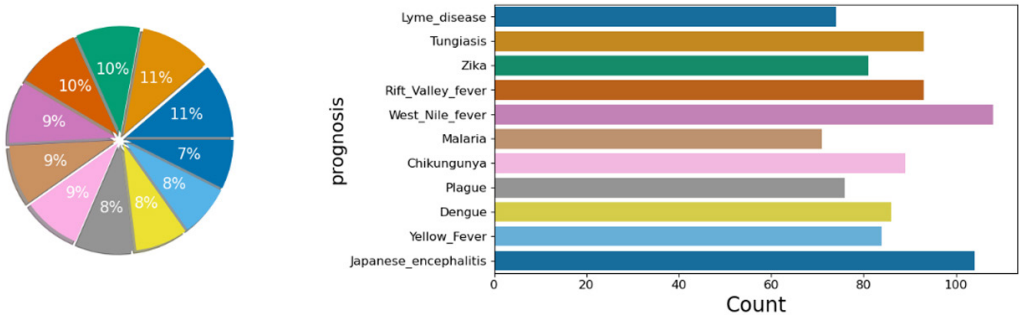

Fig. 3. Distribution of the prognosis variable

In Figure 4, the variable relationship matrix providing crucial information is presented. The interpretation of this matrix is based on the degree of correlation between different pairs of symptoms. When the matrix shows a high correlation between two symptoms, such as ulcers and nail loss, it indicates that both symptoms may be due to the same disease. On the other hand, if the matrices show a low correlation, in the case of slow heart rate and itching, the likelihood that the two symptoms are related to the same disease decreases. Then, in cases where the matrices show a normal correlation, as in stagnation and collapse, both symptoms may be related to the same disease but have different causes.

The matrix shows that people with a sudden fever often have a headache. It is also evident that people with a sudden fever tend to have nosebleeds. This could be due to dilation of the blood vessels caused by the fever, and people with headaches also tend to have nosebleeds, muscle pain, and diarrhea. The latter could be due to intestinal irritation caused by stress or anxiety accompanying the headache. People with nosebleeds tend to have a reduced range of motion. This could be due to muscle weakness caused by blood loss. People with skin rashes tend to have a decreased range of motion. This could be due to itching and irritation that may accompany the rash.

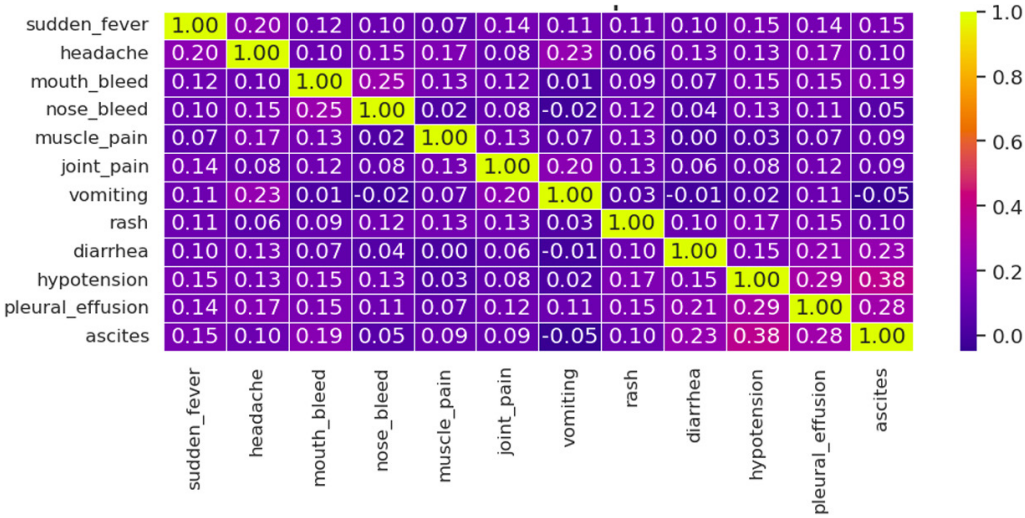

Fig. 4. Correlation of the first 10 variables

In Figure 5a, Lyme disease is scattered in the PCA chart, while tungiasis disease is clustered in the lower left part of the chart. This suggests that these diseases have different characteristics that distinguish between them, while Figure 5b shows that VBDs are more mixed than in the PCA graph. This suggests that the diseases may share some characteristics. These graphical representations provide a deeper understanding of the clustering of VBDs and the characteristics they share.

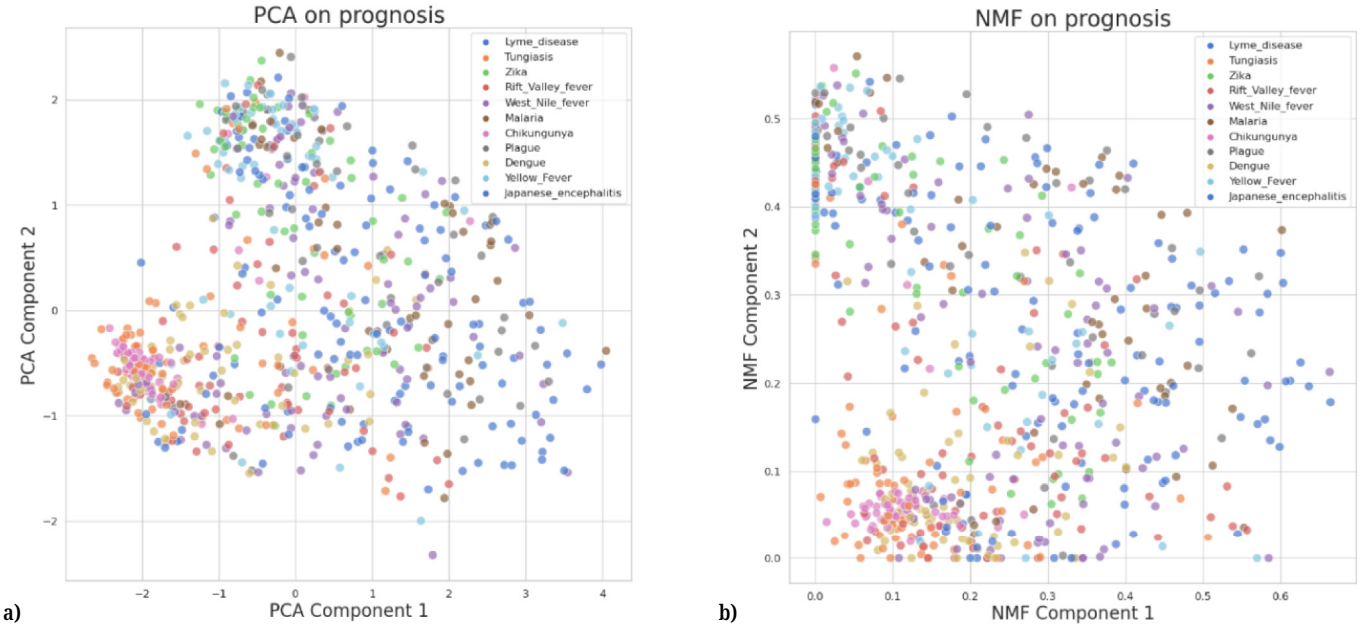

Fig. 5. Two-dimensional decomposition techniques: (a) PCA analysis, (b) NMF analysis

Data processing and modeling: Before starting the training of the data, a preprocessing phase was carried out to improve the performance of the ML models. First, the dataset was divided into training and test sets with their respective labels. Then, libraries for ML model training and evaluation, and data processing techniques were imported, such as category coding and dimensionality reduction using techniques such as PCA, NMF, UMAP, and t-SNE. Then, a class called "Decomp" was defined to perform dimensionality reduction of the data using different methods, such as PCA, NMF, UMAP, and t-SNE. These methods are used to conduct training and evaluation of datasets. It also adds the function of calculating metrics such as Logloss and MAP@3 during the training and evaluation process. Finally, the learning process is performed by cross-validation to fit the ML models. This series of tasks constitutes an integrated phase of data processing and modeling.

4 RESULTS

In this study, different ML models were investigated to predict VBDs. The models examined were SVC, XGBoost, LightGBM, CatBoost, BRF, and RF. The Ensemble Logloss score was used to evaluate the performance of each model. This metric evaluates the quality of the probabilistic predictions of a classification model. A lower Logloss value indicates better performance because there is less discrepancy between the actual data characteristics and the predicted probabilities. The average Ensemble Logloss score was 1.63822 with a standard deviation of 0.04825.

This means that the combination of the models generates acceptable predictions and that there is reasonable consistency between the various runs of the data set.

When looking at the contribution of each model to the ensemble, all models contribute to the overall accuracy in different ways. The weight assigned to each model indicates how much each contributes to the overall ensemble prediction compared to other models. In the final combination, of all the SVC models, XGBoost, LightGBM, CatBoost, BRF, and RF were significant, with weights of 0.49959 ± 0.27112, 0.58496 ± 0.22619, 0.48482 ± 0.29971, 0.54992 ± 0.27982, 0.24924 ± 0.22654 and 0.45592 ± 0.25849. With a mean of 0.24924 and a standard deviation of 0.22654, it is observed that the BRF model exhibited the lowest log loss. This suggests that the BRF performance was more consistent and contributed significantly to the overall Ensemble accuracy.

These findings show that the SVC, XGBoost, LightGBM, CatBoost, BRF, and RF models provide a robust and effective solution for predicting VBDs. The combination of several ML algorithms allows the individual strengths of each model to be leveraged, resulting in a significant improvement in predictive accuracy and the ability to address the complexities associated with VBDs prediction. These metrics can be seen in Table 2.

Table 2. Training results

Model	MAP@3 | Logloss	Accuracy	Recall	Specificity	F1-Score
SVC	0.49627 ± 0.26536	0.39309	0.38542	0.39101	0.36904
XGBoost	0.57409 ± 0.24179	0.44100	0.43229	0.44176	0.42948
LightGBM	0.46328 ± 0.30474	0.42362	0.41146	0.42417	0.40254
CatBoost	0.55200 ± 0.26964	0.46534	0.46354	0.46744	0.45971
BRF	***0.24431 ± 0.22170***	0.42341	0.42708	0.42207	0.41430
RF	0.44856 ± 0.25820	0.41445	0.42188	0.41848	0.41306

As seen in Table 2, the BRF model has the lowest MAP@3 | Logloss (0.24431 ± 0.22170) and has the lowest standard deviation, indicating that its predictions are more consistent, although consistently less accurate than the other models. Then, despite its low MAP@3 | Logloss, BRF shows comparable performance in accuracy (0.42341), recall (0.42708), specificity (0.42207), and F1-score (0.41430) concerning some other models such as RF and LightGBM.

5 DISCUSSION

Vector-borne diseases represent a serious global health threat. These disorders are spread primarily by infected insects or directly from person to person. It is estimated that they cause more than 700,000 deaths annually worldwide. In recent years, these diseases have increased worldwide. Machine learning models are essential because they can explore and process massive clinical datasets to predict diseases. Therefore, this study aims to create a predictive model using epidemiological data in tabular format to categorize and anticipate the presence of VBDs. For this purpose, the ML models, SVC, XGBoost, LightGBM, CatBoost, BRF, and RF, were applied. Regarding the Ensemble Logloss metric, the BRF model exhibits the lowest log loss, registering an average of 0.24924. According to this metric, it is concluded that this model has achieved the best performance in the ensemble. These results

differ from the study [28] where the DT model was ranked as the best predictor with an accuracy of 79%. On the other hand, the study [29] positioned the SVM model as the best predictor with an accuracy level of 76%. In this case, the predictor was based on cases reported in Jeddah, Saudi Arabia. The study [30] concluded that the LR classifier is the most accurate with 97.14%. This study focuses on separate classifiers based on symptoms and history, whereas the presented study focuses on combining different ML models for VBDs prediction. Similarly, the study [31] concluded that the LR model achieved the best performance with 83% accuracy. On the other hand, in the study [32], RF and EBM stood out in accuracy with 84%. Also, in the study [33], the RF algorithm achieved the highest accuracy of 92%. They compared the performance of the models using rapid diagnostic tests. These results support the idea that ML models can play an important role in VBDs prediction. However, these results emphasize the importance of high-quality data sets for optimal performance.

6 CONCLUSIONS

The use of machine-learning models in healthcare is constantly expanding. Therefore, it is crucial to develop a model guaranteeing effectiveness and efficiency. This study assembled multiple ML models, including SVC, XGBoost, LightGBM, CatBoost, BRF, and RF. A dataset of 1262 records was used to train these models. When training the models, it was observed that the BRF model stood out by achieving the lowest loss according to the Ensemble Logloss metric, with an average value of 0.24924 and a standard deviation of 0.22654, thus demonstrating that this model is particularly effective for VBD prediction.

We recommend using cross-validation techniques to evaluate performance more accurately and optimizing hyperparameters to enhance accuracy. Finally, the models used in this study showed promising results, which could contribute significantly to the early detection of patients with vector-borne diseases.

The results obtained in this study show that combining the SVC, XGBoost, LightGBM, CatBoost, BRF, and RF models forms a robust and effective ensemble for VBD prediction. The different approaches of these models worked together to improve overall performance. Each model made a unique contribution to the ensemble, leveraging individual strengths to enhance the ensemble's performance.

7 REFERENCES

[1] M. Arquam, A. Singh, and H. Cherifi, "Impact of seasonal conditions on vector-borne epidemiological dynamics," *IEEE Access*, vol. 8, pp. 94510–94525, 2020. https://doi.org/10.1109/ACCESS.2020.2995650

[2] T. D. Hollingsworth, J. R. C. Pulliam, S. Funk, J. E. Truscott, V. Isham, and A. L. Lloyd, "Seven challenges for modeling indirect transmission: Vector-borne diseases, macroparasites and neglected tropical diseases," *Epidemics*, vol. 10, pp. 16–20, 2015. https://doi.org/10.1016/j.epidem.2014.08.007

[3] A. Otten, A. Fazil, A. Chemeris, P. Breadner, and V. Ng, "Prioritization of vector-borne diseases in Canada under current climate and projected climate change," *Microbial Risk Analysis*, vol. 14, p. 100089, 2020. https://doi.org/10.1016/j.mran.2019.100089

[4] A. M. George, R. Ansumana, D. K. de Souza, V. K. M. Niyas, A. Zumla, and M. J. Bockarie, "Climate change and the rising incidence of vector-borne diseases globally," *International Journal of Infectious Diseases*, vol. 139, pp. 143–145, 2024. https://doi.org/10.1016/j.ijid.2023.12.004

[5] C. Davitt, R. Traub, B. Batsukh, B. Battur, M. Pfeffer, and A. K. Wiethoelter, "Knowledge of Mongolian veterinarians towards canine vector-borne diseases," *One Health*, vol. 15, p. 100458, 2022. https://doi.org/10.1016/j.onehlt.2022.100458

[6] S. Hussain *et al.*, "First molecular confirmation of multiple zoonotic vector-borne diseases in pet dogs and cats of Hong Kong SAR," *Ticks and Tick-borne Diseases*, vol. 14, no. 4, p. 102191, 2023. https://doi.org/10.1016/j.ttbdis.2023.102191

[7] O. Saucedo and J. H. Tien, "Host movement, transmission hot spots, and vector-borne disease dynamics on spatial networks," *Infectious Disease Modelling*, vol. 7, no. 4, pp. 742–760, 2022. https://doi.org/10.1016/j.idm.2022.10.006

[8] Z. S. Rahmat, M. Sadiq, L. I. Vohra, H. Ullah, and M. Y. Essar, "The impact of COVID-19 followed by extreme flooding on vector-borne diseases in Pakistan: A mini-narrative review," *New Microbes and New Infections*, vol. 51, p. 101075, 2023. https://doi.org/10.1016/j.nmni.2022.101075

[9] Y. H. Kao and M. C. Eisenberg, "Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment," *Epidemics*, vol. 25, pp. 89–100, 2018. https://doi.org/10.1016/j.epidem.2018.05.010

[10] S. Schorderet-Weber, S. Noack, P. M. Selzer, and R. Kaminsky, "Blocking transmission of vector-borne diseases," *International Journal for Parasitology: Drugs and Drug Resistance*, vol. 7, no. 1, pp. 90–109, 2017. https://doi.org/10.1016/j.ijpddr.2017.01.004

[11] T. A. Taz, M. Kawsar, B. K. Paul, K. Ahmed, and T. Bhuyian, "Characterizing topological properties and network pathway model among vector-borne diseases," *Informatics in Medicine Unlocked*, vol. 18, p. 100312, 2020. https://doi.org/10.1016/j.imu.2020.100312

[12] B. Karuppusamy, D. K. Sarma, P. Lalmalsawma, L. Pautu, K. Karmodiya, and P. Balabaskaran Nina, "Effect of climate change and deforestation on vector-borne diseases in the North-Eastern Indian State of Mizoram bordering Myanmar," *The Journal of Climate Change and Health*, vol. 2, p. 100015, 2021. https://doi.org/10.1016/j.joclim.2021.100015

[13] I. Zortman *et al.*, "A social-ecological systems approach to tick bite and tick-borne disease risk management: Exploring collective action in the Occitanie region in southern France," *One Health*, vol. 17, p. 100630, 2023. https://doi.org/10.1016/j.onehlt.2023.100630

[14] C. Olechnowicz, J. Leahy, A. Gardner, and C. C. Sponarski, "Perceived vulnerability for Lyme disease questionnaire: A social science tool for understanding tick-borne disease attitudes," *Ticks and Tick-borne Disease*, vol. 14, no. 2, p. 102120, 2023. https://doi.org/10.1016/j.ttbdis.2023.102120

[15] M. Ferraguti, A. Dimas Martins, and Y. Artzy-Randrup, "Quantifying the invasion risk of West Nile virus: Insights from a multi-vector and multi-host SEIR model," *One Health*, vol. 17, p. 100638, 2023. https://doi.org/10.1016/j.onehlt.2023.100638

[16] C. Pley, M. Evans, R. Lowe, H. Montgomery, and S. Yacoub, "Digital and technological innovation in vector-borne disease surveillance to predict, detect, and control climate-driven outbreaks," *Lancet Planet Health*, vol. 5, no. 10, pp. E739–E745, 2021. https://doi.org/10.1016/S2542-5196(21)00141-8

[17] F. Dantas-Torres and D. Otranto, "Best practices for preventing vector-borne diseases in dogs and humans," *Trends Parasitology*, vol. 32, no. 1, pp. 43–55, 2016. https://doi.org/10.1016/j.pt.2015.09.004

[18] Y. A. Alqudah, B. Sababha, E. Qaralleh, and T. Youssef, "Machine learning to classify driving events using mobile phone sensors data," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 2, pp. 124–136, 2021. https://doi.org/10.3991/ijim.v15i02.18303

[19] H. Guliyev, "Artificial intelligence and unemployment in high-tech developed countries: New insights from the dynamic panel data model," *Research in Globalization*, vol. 7, p. 100140, 2023. https://doi.org/10.1016/j.resglo.2023.100140

[20] F. Guo and H. Meng, "Application of artificial intelligence in gastrointestinal endoscopy," *Arab Journal of Gastroenterology*, vol. 25, no. 2, pp. 92–96, 2024. https://doi.org/10.1016/j.ajg.2023.12.010

[21] A. D. Rebelo, D. E. Verboom, N. R. dos Santos, and J. W. de Graaf, "The impact of artificial intelligence on the tasks of mental healthcare workers: A scoping review," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100008, 2023. https://doi.org/10.1016/j.chbah.2023.100008

[22] P. Moingeon, "Artificial intelligence-driven drug development against autoimmune diseases," *Trends Pharmacol Sci*, vol. 44, no. 7, pp. 411–424, 2023. https://doi.org/10.1016/j.tips.2023.04.005

[23] A. V. Singh *et al.*, "Integrative toxicogenomics: Advancing precision medicine and toxicology through artificial intelligence and OMICs technology," *Biomedicine & Pharmacotherapy*, vol. 163, p. 114784, 2023. https://doi.org/10.1016/j.biopha.2023.114784

[24] P. N. Jone *et al.*, "Artificial intelligence in congenital heart disease: Current state and prospects," *JACC: Advances*, vol. 1, no. 5, p. 100153, 2022. https://doi.org/10.1016/j.jacadv.2022.100153

[25] H. E. Bays *et al.*, "Artificial intelligence and obesity management: An Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2023," *Obesity Pillars*, vol. 6, p. 100065, 2023. https://doi.org/10.1016/j.obpill.2023.100065

[26] T. Hariguna and A. Ruangkanjanases, "Assessing the impact of artificial intelligence on customer performance: A quantitative study using partial least squares methodology," *Data Science and Management*, vol. 7, no. 3, pp. 155–163, 2024. https://doi.org/10.1016/j.dsm.2024.01.001

[27] M. R. Al Nasar, I. Nasir, T. Mohamed, N. S. Elmitwally, M. M. Al-Sakhnini, and T. Asgher, "Detection of dengue disease empowered with fused machine learning," *International Conference on Cyber Resilience (ICCR)*, 2022, pp. 1–10. https://doi.org/10.1109/ICCR56254.2022.9996009

[28] D. Sarma, S. Hossain, T. Mittra, M. A. M. Bhuiya, I. Saha, and R. Chakma, "Dengue prediction using machine learning algorithms," in *IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2020, pp. 1–6. https://doi.org/10.1109/R10-HTC49770.2020.9357035

[29] A. Siddiq, N. Shukla, and B. Pradhan, "Predicting dengue fever transmission using machine learning methods," in *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2021, pp. 21–26, 2021. https://doi.org/10.1109/IEEM50564.2021.9672977

[30] Y. P. Bria, C. H. Yeh, and S. Bedingfield, "Machine learning classifiers for symptom-based malaria prediction," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–6. https://doi.org/10.1109/IJCNN55064.2022.9891945

[31] E. Mbunge, R. C. Millham, M. N. Sibiya, and S. Takavarasha, "Application of machine learning models to predict malaria using malaria cases and environmental risk factors," in *2022 Conference on Information Communications Technology and Society, (ICTAS)*, 2022, pp. 1–5. https://doi.org/10.1109/ICTAS53252.2022.9744657

[32] S. Rajab, J. Nakatumba-Nabende, and G. Marvin, "Interpretable machine learning models for predicting malaria," in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing, (ICSTSN)*, 2023, pp. 1–6. https://doi.org/10.1109/ICSTSN57873.2023.10151538

[33] S. S. Yadav, V. J. Kadam, S. M. Jadhav, S. Jagtap, and P. R. Pathak, "Machine learning based malaria prediction using clinical findings," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, pp. 216–222. https://doi.org/10.1109/ESCI50559.2021.9396850

[34] N. Shabanpour, S. V. Razavi-Termeh, A. Sadeghi-Niaraki, S. M. Choi, and T. Abuhmed, "Integration of machine learning algorithms and GIS-based approaches to cutaneous leishmaniasis prevalence risk mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102854, 2022. https://doi.org/10.1016/j.jag.2022.102854

[35] H. Moosaei and M. Hladík, "Sparse solution of least-squares twin multi-class support vector machine using ℓ0 and ℓp-norm for classification and feature selection," *Neural Networks*, vol. 166, pp. 471–486, 2023. https://doi.org/10.1016/j.neunet.2023.07.039

[36] Q. Si, Z. Yang, and J. Ye, "Symmetric LINEX loss twin support vector machine for robust classification and its fast iterative algorithm," *Neural Networks*, vol. 168, pp. 143–160, 2023. https://doi.org/10.1016/j.neunet.2023.08.055

[37] M. Almseidin, A. M. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 13, no. 12, pp. 171–183, 2019. https://doi.org/10.3991/ijim.v13i12.11411

[38] J. Zheng *et al.*, "Metabolic syndrome prediction model using Bayesian optimization and XGBoost based on traditional Chinese medicine features," *Heliyon*, vol. 9, no. 12, 2023. https://doi.org/10.1016/j.heliyon.2023.e22727

[39] Y. Song *et al.*, "Spatial prediction of PM2.5 concentration using hyper-parameter optimization XGBoost model in China," *Environmental Technology & Innovation*, vol. 32, p. 103272, 2023. https://doi.org/10.1016/j.eti.2023.103272

[40] Z. Pan, W. Lu, and Y. Bai, "Groundwater contaminated source estimation based on adaptive correction iterative ensemble smoother with an auto lightgbm surrogate," *Journal of Hydrology*, vol. 620, p. 129502, 2023. https://doi.org/10.1016/j.jhydrol.2023.129502

[41] B. Li *et al.*, "High-spatiotemporal-resolution dynamic water monitoring using LightGBM model and Sentinel-2 MSI data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103278, 2023. https://doi.org/10.1016/j.jag.2023.103278

[42] S. Chehreh Chelgani, H. Nasiri, A. Tohry, and H. R. Heidari, "Modeling industrial hydrocyclone operational variables by SHAP-CatBoost – A 'conscious lab' approach," *Powder Technol.*, vol. 420, p. 118416, 2023. https://doi.org/10.1016/j.powtec.2023.118416

[43] S. Zhang, X. Lu, and Z. Lu, "Improved CNN-based CatBoost model for license plate remote sensing image classification," *Signal Processing*, vol. 213, p. 109196, 2023. https://doi.org/10.1016/j.sigpro.2023.109196

[44] Ł. Kobyliński and A. Przepiórkowski, "Definition extraction with balanced random forests," in *Advances in Natural Language Processing, GoTAL 2008,* in Lecture Notes in Computer Science, B. Nordström and A. Ranta, Eds., Springer, Berlin, Heidelberg, vol. 5221, 2008, pp. 237–247. https://doi.org/10.1007/978-3-540-85287-2_23

[45] R. Achawanantakun, J. Chen, Y. Sun, and Y. Zhang, "LncRNA-ID: Long non-coding RNA IDentification using balanced random forests," *Bioinformatics*, vol. 31, no. 24, pp. 3897–3905, 2015. https://doi.org/10.1093/bioinformatics/btv480

[46] O. Iparraguirre-Villanueva *et al.*, "Comparison of predictive machine learning models to predict the level of adaptability of students in online education," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 4, 2023. https://doi.org/10.14569/IJACSA.2023.0140455

[47] W. Zhou, Z. Liang, Z. Fan, and Z. Li, "Spatio–temporal effects of the built environment on running activity based on a random forest approach in Nanjing, China," *Health Place*, vol. 85, p. 103176, 2024. https://doi.org/10.1016/j.healthplace.2024.103176

[48] O. Iparraguirre-Villanueva, A. Epifanía-Huerta, C. Torres-Ceclén, J. Ruiz-Alvarado, and M. Cabanillas-Carbonell, "Breast cancer prediction using machine learning models," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 2, 2023. https://doi.org/10.14569/IJACSA.2023.0140272

8 AUTHORS

Orlando Iparraguirre-Villanueva is a systems engineer with a master's degree in information technology management and a Ph.D. in systems engineering from Universidad Nacional Federico Villarreal-Peru. ITIL® Foundation Certificate in IT Service, Specialization in Business Continuity Management, Scrum Fundamentals Certification (SFC). National and international speaker/panelist (Panamá, Colombia, Ecuador, Venezuela, and México) (E-mail: oiparraguirre@ieee.org).

Michael Cabanillas-Carbonell is an engineer, and has a master's in systems engineering from the National University of Callao, Peru; and a PhD candidate in systems engineering and telecommunications at the Polytechnic University of Madrid. Besides being President of the chapter of the Education Society IEEE-Peru. Conference Chair of the Engineering International Research Conference IEEE Peru EIRCON, is also a Research Professor at Norbert Wiener University, Professor at Universidad Privada del Norte, Universidad Autónoma del Perú. Advisor and Jury of Engineering Thesis in different universities in Peru. Being an international lecturer in Spain, the United Kingdom, South Africa, Romania, Argentina, Chile, and China. Specialization in Software Development, Artificial Intelligence, Machine Learning, Business intelligence, Augmented Reality, is also a Reviewer IEEE Peru and author of more than 50 scientific articles indexed in IEEE Xplore and Scopus (E-mail: mcabanillas@ieee.org).