International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 5, October, 2023, pp. 1281-1290
Journal homepage: http://iieta.org/journals/ria

Exploring Machine Learning Algorithms for the Prediction of Dengue: A Comprehensive Review

Archana Thirugnanam*, Faritha Banu Jahir Hussain

Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Bharathi Salai, Ramapuram, Chennai 600089, Tamil Nadu, India

Corresponding Author Email: archanat@srmist.edu.in

https://doi.org/10.18280/ria.370521

Received: 21 July 2023
Revised: 3 September 2023
Accepted: 9 September 2023
Available online: 31 October 2023

Keywords:
vector borne diseases, classifier, machine learning, dengue, prediction model

ABSTRACT

Vector-borne diseases, transmitted by blood-feeding arthropods like mosquitoes, ticks, and fleas, pose an escalating challenge to global public health. Dengue, a disease propagated by Aedes mosquitoes, is currently the most rapidly spreading vector-borne illness worldwide. Given its endemic nature, the prevention and control of outbreaks remain a global imperative. Timely detection of dengue is critical to mitigate mortality rates, making predictive models indispensable tools for public health planning, resource allocation, and disease control. This study undertakes a comprehensive review of various machine learning algorithms used in developing predictive models for early-stage dengue detection based on presented symptoms. The review encompasses the entire modeling process, including data preprocessing, algorithm implementation, evaluation, and validation. It further delves into the algorithms' ability to accurately classify dengue into febrile, critical, or convalescent phases. An array of machine learning algorithms, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine (SVM), Decision Tree, Artificial Neural Network, and Naive Bayes Classifier were analyzed. The advantages and disadvantages of these algorithms are discussed to identify the most effective approach for dengue prediction. The Naive Bayes algorithm was found to quickly generate predictions with a precision value of 99.1%. However, the SVM model outperformed all others with a cross-validation score of 98.5%, K-Fold validation of 97.5%, precision of 98.2%, and an F1 Score of 98.0%, thereby enhancing the overall performance of the predictive model.

1. INTRODUCTION

Vector-borne diseases are spread by blood-feeding arthropods such as mosquitoes, ticks, and fleas that cause serious illness in humans. Vector-borne diseases such as Malaria, dengue, West Nile fever, Chikungunya, Rift Valley fever, Japanese encephalitis, Yellow fever, and Zika cause around one billion critical cases and one million fatal cases every year. The impact of dengue disease goes far beyond tropical areas and places with limited resources. Aedes mosquitoes are the main vectors of the virus, which has spread quickly as a result of factors like increased international travel, urbanisation, and climate change.

There are four distinct serotypes of the dengue virus (DENV-1, DENV-2, DENV-3, and DENV-4), and infection with one serotype provides lifelong immunity against that specific serotype but only temporary immunity against the others. In some cases, subsequent infection with a different serotype can lead to more severe forms of the disease, such as dengue haemorrhagic fever (DHF) or dengue shock syndrome (DSS) [1]. When an infected mosquito bites a human, it injects the dengue virus along with its saliva into the person's bloodstream. The virus starts to replicate in various cell types, including immune cells and endothelial cells lining blood vessels. The virus replicates in the human host, leading to a period of viremia, during which the virus circulates in the bloodstream. This is when the virus can be detected and transmitted to other mosquitoes if they bite the infected person.

There are 96 million symptomatic instances and 40,000 fatalities annually as a result of it, which affects more than 3.9 billion people in 129 countries. Severe dengue is a cause of serious illness and mortality in children under the age of five. dengue fever has become endemic; therefore, preventing and controlling outbreaks of it remains a serious issue around the world. Due to the wide range of geographical and socio-economic conditions, controlling many diseases at an early stage is challenging in the current system [2].

Dengue prediction and prevention present special challenges requiring novel approaches. One major challenge lies in the complex interplay of factors influencing the disease's transmission. The patterns of dengue transmission are intimately shaped by climate factors, mosquito behaviour, population migration, and immunity dynamics. The several dengue virus strains also provide an additional level of complication. The impact of interactions between various strains on the severity of the illness and immunity could have an impact on the efficiency of treatments and the precision of forecasts. This demands adaptable models that can take into account viral variety and evolution. Furthermore, the absence of a definitive vaccine or cure for dengue heightens the significance of prediction. Therefore, predictive modelling of dengue acquires essential significance, providing an approach to reduce its catastrophic impacts, ultimately leading to better health outcomes for communities at risk.

Predicting dengue outbreaks is crucial for public health management, disease prevention, and reducing the impact of this potentially severe illness on individuals, communities, and economies. Furthermore, the disease's potential to develop into serious forms like dengue haemorrhagic fever and dengue shock syndrome highlights the need for quick development of preventative measures.

The paper aims to study different Machine Learning algorithms and find the effective algorithm that can assist the medical practitioner for early prediction and diagnosis of dengue. The scope revolves around utilizing these algorithms to enhance dengue diagnosis and patient case.

The main objective is to analyse the various machine learning algorithms to propose a dengue predictive model is that it can contribute to early detection, prevention, and control of dengue outbreaks. Creating an early warning system based on the predictive model also allows public health authorities to anticipate dengue before they escalate. This proactive approach can help prevent the severity of the disease and reduce its impact. Predicting dengue can help healthcare systems be prepared for possible outbreaks. Hospitals and healthcare providers can allocate resources and manpower based on the predicted disease.

This study helps to identify algorithms to develop a dengue prediction model. The majority of the prediction model consists of three modules: pre-processing, classification, and Validation and testing. Figure 1 shows the data preparation stages.

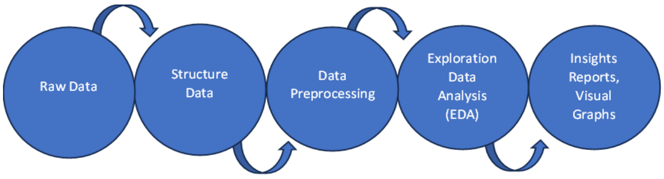
Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insights Reports, Visual Graphs

Figure 1. Data preparation stages

The proposed study compares the performance of different machine learning algorithms, namely Support Vector Machine (SVM), Random Forest Classifier, Artificial Neural Network, Ensemble Technique, Naive Bayes, and k-nearest neighbour classifier for the prediction of dengue. Comparing these algorithms contributes to the prediction of dengue by helping to select the best-performing models, understanding feature importance, improving generalization, exploring ensemble approaches, fine-tuning hyperparameters, ensuring interpretability, optimizing resource usage, handling data challenges, and maintaining model relevance over time.

The remainder of the paper is structured as follows: the second section covers the background and related works of the current machine learning algorithms while the third section elaborates the experimental results. The fourth section gives the conclusion of the study.

2. BACKGROUND AND RELATED WORK

A branch of AI that has existed for many years is Machine Learning. In the last few years, the discipline has rapidly expanded and new methods are being created every day. The four categories of ML health therapies are patient diagnostics, mortality risk assessment or patient morbidity, health management planning, and infectious disease outbreak prediction and monitoring. The diagnosis procedure is improved by sophisticated AI approaches, which increase its accuracy and dependability. Due to their capacity for illness detection, ML methods [3] have become a vital and essential component of the diagnosis sector. The amount of health data now available has increased dramatically. The system uses secure storage and access for all data [4]. Huge amounts of data on healthcare are exchanged and processed cooperatively by members of the big data community. The community's goal is to apply the most up-to-date data-processing and analytic techniques to enhance the effectiveness and quality of healthcare.

One of the sectors in the global economy with the most data is the healthcare sector. In the current situation, patients and the healthcare providers provide an enormous amount of data. Healthcare delivery is significantly improved by data collection, organization, and analysis; however, the missing values, the specific reason for the missing values, the validity of data model, the requirement for several forms of data and the operational viability are the key issues faced by the data models of Electronic Health Records (EHR) [5, 6]. Pandey and Janghel [7] have reviewed various ML methods for forecasting the beginning of illnesses using EHR data. The researchers emphasised the practise of testing a model using a variety of machine-learning approaches and utilising strategies like strong feature selection, data size for enhancing the models for standard clinical procedures in hospital environments. The study emphasises how several machine learning approaches have been successful in foretelling viral illnesses [8, 9].

2.1 Logistic regression

One or more categorical explanatory factors are modelled against a continuous response variable using the statistical approach known as Logistic Regression (LR). The independent variable (x) is transformed by the sigmoid function into a probability expression between 0 to 1.

Logistic regression is a statistical method used for predicting the probability of a binary outcome (i.e., an outcome that can be classified as either "success" or "failure"). In the case of dengue prediction, logistic regression can be utilized to forecast the likelihood of a person contracting dengue fever based on various risk factors. The age of the individual, gender, socioeconomic position, travel history (recently visited region with a high prevalence of dengue), Prior dengue infection history, the presence of mosquito breeding grounds closes to the person's place of residence, and other additional characteristics could all serve as variables that predict dengue in a logistic regression model.

Once the data are collected and the variables predicted, Python or other statistical software package is used to implement a logistic regression model. The output of the model is used to calculate the predicted probability of dengue infection for each individual in the dataset.

Caicedo et al. developed a prediction model using machine learning approaches for forecasting dengue and chikungunya in Colombia. This study's main objective is to use time series to develop a prediction model for the dynamics of dengue and chikungunya disease. The author has used three machine learning algorithms—Kernel Regression, Gaussian processes, and time series—to create the suggested model. The study's findings suggest that Kernel Ridge Regression can be used to monitor Colombia's public health for both dengue and chikungunya. It has the best performance for forecasting dengue cases [10].

To predict the dengue epidemic, a prediction algorithm was suggested that utilises time series as data inputs from several cities and functions as a multivariate time series predictor. The Lasso regression, LSTM and Random Forest machine learning algorithms have been created and compared to this prediction model. The experimental findings showed that the LSTM outperformed both the Random Forest Classifier and the Lasso regression. Using the results of this proposed model, Brazilian cities should be predicted [11].

2.2. K-nearest neighbour

K-nearest neighbour (KNN) algorithm can be applied for the prediction of dengue. KNN is a non-parametric algorithm, which implies that it does not make any assumptions about the distribution of the underlying data. Instead, it makes predictions depends on the data that is found to be close to the new data point and then it is evaluated. The dataset for Dengue prediction using KNN is similar to those used in logistic regression, such as age, gender, travel history, and presence of mosquito breeding sites. In addition, KNN can also take into account variables that may not have a linear relationship, such as distance to the nearest hospital or the number of people living in the household.

To use KNN for dengue prediction, the dataset must be divided into two sets namely training sets and testing sets. The KNN model must then be trained with the help of training set, and the effectiveness of the training set must be evaluated based on the testing set. The KNN algorithm finds the K data points that are closest to the new data point being evaluated in the training set (i.e., the "nearest neighbours"), and using the outcome variable values for those neighbours to predict the outcome for the new data point. The value of K in KNN is an important hyperparameter that has to be chosen carefully. In situations where the genuine decision boundary is substantially non-linear, a greater value of K will produce a smoother decision border that results in worse performance [12].

Using the number (K) of neighbours, K-nearest neighbour (KNN) algorithm can perform classification and regression problems. The predicted value is determined by selecting the k observations that are most comparable to each query point [13]. A variety of disease diagnosis models are developed using the KNN algorithm. These factors were taken as reference for the prediction of disease. Vijayakumar et al. presents a prediction model that, utilising demographic information acquired from web applications and mobile applications, may identify vector-borne disease. This approach can also determine the similarity across diseases if a patient displays symptoms that are similar to those of any vector-borne illness [14]. The author used the fuzzy K-nearest neighbour classifier to assess whether a patient was infected or not. Additionally, the author has developed a smart system that uses IoT sensors to locate mosquito breeding grounds and densities in different geographical areas. In order to pinpoint dangerous areas, a social network graph has been built using the prediction model. This model can deliver registration alert messages to users who have registered to avoid disease outbreak. The finding of this paper is that this proposed model has achieved 95.9% classification accuracy.

2.3 Support vector machine

In order to handle both linear and non-linear data, Support Vector Machine (SVM) finds a hyperplane decision boundary that maximum separates the two classes in the data [15]. SVM develops a representation of all the data points which divides or segregates distinct labels/categories into as many as feasible distinct gaps. The most effective hyperplane for dividing the clusters/classes is discovered by the method. Support vectors for the hyperplane can be found nearby [16]. Least squares SVM, successive projection algorithm-SVM and support vector regression are the three widely used SVM methods which comprise Pattern recognition and classification are two areas where SVMs are heavily used, and they have also been successfully applied to a number of real-world issues [17, 18].

Nordin et al. [19] has proposed a prediction model for the classification of dengue endemic cases using Support Vector Machine. The author has gathered data samples of dengue patients from the Health Department of Kelantan, Malaysia. The authors identified several symptoms that were strongly associated with dengue fever such as fever, headache and joint pain, which were included in the logistic regression model.

A prediction model predicting morbidity rate of dengue cases in Thailand [20] was predicted using K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT). According to this study's conclusion, a rise in the number of female mosquitoes and their larvae in a particular area was identified. The author has applied support vector machine the radial using basis function and achieved 88.37 % prediction accuracy of proposed model.

Algorithm 1. SVM Classification Algorithm to evaluate the accuracy of dengue Prediction model

```
1: Import the necessary Libraries
2: Load the dataset in a CSV file
3: Split the dataset into Training set and testing set (X_train, X_test, y_train, y_test)
4: Set the Predictor variable (x)
x=data.drop('has_dengue', axis=1)
5: Set the Target variable (y)
y=data['has_dengue']
6: Create the model for SVM
7: Train the SVM model on the training set
8: Predict the target variable for the testing set (y_pred)
9: Evaluate the accuracy of the model
accuracy=accuracy_score(y_test, y_pred)
10: Print the accuracy
11: end
```

A novel biological finding made using nontrivial data mining utilising currently available computer methods. The system's objective is to facilitate the gathering and retrieval of public health-related papers, data, learning resources, and tools. In addition to facilitate data integration and information sharing in the area of dengue, one of the most common viral infections have established this general infrastructure. The SVM classification with the radial basis function was introduced in this study as a method for categorising viral data, and the model shows a very accurate prediction rate [21].

In the algorithm 1, the required libraries for SVM are imported and then the dataset from a CSV file are loaded using the read_csv function from Pandas. The function train_test_split from Scikit-Learn is then used to split the dataset into training set and testing set. Next, an SVM model using the SVC class from Scikit-learn are created, specifying a linear kernel. The model object's fit method is used to train the SVM model on the given training set. In this algorithm, the trained model is utilized to forecast the target variable for the testing set. The function accuracy_score from Scikit-learn is used to relate the predicted values with the actual values in the testing set in order to evaluate the model's accuracy.

To optimize the model's performance, the dataset must be checked whether it is appropriately prepared for the SVM algorithm, including handling missing data, scaling the predictor variables, and dealing with any imbalanced classes in the target variable. Dourjoy et al. have proposed a comparative analysis of dengue fever prediction using machine learning algorithms which aimed to develop an SVM-based system for initial prediction of dengue fever by using symptoms reported by the patients [22]. The study used a dataset of patients with suspected dengue fever. The dataset included information on various symptoms such as fever, headache, joint pain, and nausea. The authors used an SVM algorithm to predict whether a patient had dengue fever based on the presence of specific symptoms. It also sought to incorporate other parameters associated to the disease that are required to improve the model's ability to accurately predict outcomes when used in regions with similar climatic conditions and when weather variables like temperature, humidity and total rainfall are not significantly different. Their model described that the SVM-based system was able to detect dengue fever with an accuracy of 94.4%. The authors identified several symptoms that were strongly associated with dengue fever such as fever, headache and rash which were included in the SVM model.

2.4 Decision tree

A Supervised learning technique known as Decision Tree (DT) makes predictions and decisions by learning from the previous history of data. It splits the dataset as two major groups based on the attribute value or random value and then it compares both the groups to find the best split point. Few approaches like Classification and regression trees (CART), chi-squared automatic interaction detection (CHAID) and Iterative dichotomiser (ID) are used to construct a DT. DTs are employed in decision-making, object classification, result prediction, and data analysis. DTs are used in medicine to find new drugs and diagnose diseases [23-25].

Several researchers have used decision tree algorithms for early prediction of dengue fever from symptoms for patients on the base of geographical locations. Kalansuriya et al. [26] aimed to build a model for location-based dengue prediction of dengue fever based on symptoms reported by patients. A methodology has been put out by Pravin et al. to identify dengue patients at an early stage. They have retrieved patient data from suspected dengue patients using IoT sensors in the framework. This framework has classified people records on the basis of symptoms [27].

Rao et al developed a prediction model in real time which is used to identify dengue at the initial stage. It also minimizes the number of false negative and false positive results of the prediction model. It is then trained and tested with the datasets. The implementation of decision tree on prediction model has achieved highest accuracy [28]. The propose model can be used as a universal tool for dengue diagnosis.

2.5 Naive bayes

A probabilistic classifier namely Naive Bayes (NB) is modified with some simplifications that is built on the Bayes theorem. The features in the NB classifier are presumably conditionally independent based on the label. NB has been used for weather forecasting, spam filtering, and medical diagnosis [29, 30].

Naive Bayes classifiers are built on feature selection algorithms, which determine the conditional likelihood of features belonging to a class, which are frequently used for text classification in machine learning [31]. After determining features using a recognised feature selection approach, an auxiliary feature technique is utilised to choose an auxiliary feature that can reclassify the text space targeted at the supplied features. The appropriate conditional probability is then altered to improve classification accuracy.

Harumy et al. [32] discussed about the different factors responsible for spreading dengue in the domain of his country by using Artificial Neural Network (ANN), Naïve Bayes, Backpropagation, Regression. The proposed model has three layers. In the first layer, data collection and identification of variables are performed. While second phase is of comparison of prediction of dengue cases. Third phase identifies factors that most influences. The outcome reveals that the proposed prediction model has an accuracy rate of 87.16%.

Peng et al. [33] proposed a prediction model using time series which was active to dengue and chikungunya. The following methods namely Gaussian processes, Kernel Ridge Regression are used in the findings. The study concludes that the Kernel Ridge Regression method has the maximum performance for predicting dengue. It is also capable of modelling linear and nonlinear relationships between predictor variables and outcomes.

During the period 2013 to 2017 a survey was conducted in five areas in and around Malaysia which saw the greatest incidence of dengue sickness. The survey evaluated the most reliable machine learning model for predicting dengue outbreaks [34]. The factors that affect the climate like temperature, humidity, wind speed and rainfall were included in every model. According to the data, the SVM showed the best capacity to predict with an accuracy of 70%, sensitivity of 14%, specificity of 95% and precision of 56%. However, with reference to the imbalanced data (original data), the testing sample's SVM (linear) sensitivity increased to 63.54% from 14.4%. The SVM model's most significant predictor was the week of the year. This study serves as an example of the potential machine learning for the forecasting of dengue outbreaks.

The datasets include symptoms such as blood pressure, fever, heart rate, temperature and platelets. It also includes symptoms that both the patient and the doctor notice. The use of the Naive Bayes classification algorithm for predicting dengue sickness is described in the consultation. Real-time data is accepted and saved on the server through the analysis of the patient's earlier medical information [35]. Additionally, the dataset was compared to the acceptable real-time data, which provides the likelihood that dengue will occur.

A machine learning-based classification algorithm that uses a patient's likelihood of being admitted to the paediatric intensive care unit as a surrogate for the severity of dengue. A huge number of machine learning techniques were well trained and validated with the help of Stratified 5-Fold cross-validation. The finest model was chosen and then it was evaluated on a separate test set. The findings of the Cross-validation reveal that SVM with a Gaussian Kernel beat the other existing models and with Area under the ROC curve (AUC ROC) achieve a score of 0.81. Later results across the test set displayed an average AUC ROC scores a value of 0.75. The findings of validation and testing are encouraging and also encourage additional study and advancement [36].

2.6 Neural networks

Neural networks (NN) are a collection of layers of neurons and are modelled after the organisation of the human brain. Input, hidden, and output layers are how these neurons are organised. Input x_i, weights w_i, and bias (b) are collected as a numerical value. The input to the neuron Sum is determined as defined in Eq. (1).

$$\text{Sum}=\sum_{i=1}^{n} x_i w_i + b \qquad (1)$$

The neurons in the buried layers of the brain receive this Sum value as input. The threshold function t uses the estimated value to activate a neuron. Some known standard neural networks are Artificial Neural Network (ANN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). A number of neurons are coupled by back-propagation and feed-forward forms in ANN [37, 38]. CNN employs a multi-layer perceptron variant [39]. ANN models rely heavily on hardware but the CNN model requires a large amount of training data to function effectively and does not record the object's location or orientation. Voice, handwriting, and picture recognition are all visual patterns that these networks can be trained to recognise [40, 41]. Modern NNs are hence much more effective than earlier versions.

Gambhir et al. [42] has developed a model for an early detection of dengue disease taking a dataset of 110 patients in Delhi from different hospitals during 2015-16. The dataset is further classified in to two different classes namely dengue positive and dengue negative. In this study, the dataset is composed of 85 patients as Dengue Positive and rest as Negative. The author proposed model has been divided into three phases. The first phase of propose prediction model has performed data collection and pre-processing work. In the second phase, the K-cross fold validation method has performed for reducing unwanted data. Further machine learning technologies specifically Decision Tree (DT), Artificial Neural Network (ANN) and Naïve Bayes (NB) are used in the proposed model. This K-Cross Fold validation method validates results produced by three machine learning algorithms. Researcher has found that ANN based diagnostic model has given better performance in the study than the Naïve Bayes (NB) and Decision Tree (DT). In third phase, the proposed model is capable of taking decision whether a patient is found dengue positive or dengue Negative. The results shows that ANN -based prediction model has achieved 79 % accuracy, 55.55 % sensitivity, 88.5 % Specificity. On other hand, Naïve bayes -based prediction model has achieved 76.36 % accuracy, 47.66 % sensitivity, 84.88 % Specificity while Decision Tree -based prediction model has achieved 73.63 % accuracy, 41.66 % sensitivity, 82.55 % Specificity. The study's conclusion is that an ANN-based prediction model is ideal for identifying dengue patients at an early stage.

The public health sector's absence of a robust medical infrastructure has a particularly negative impact on rural communities. As members of the lower middle class or those who fall below the poverty line, a significant portion of the population cannot afford private hospitals. As a result, the burden on the public health sector has multiplied. In the event of an epidemic, the mortality rate is significant since the majority of the populace lacks access to healthcare and prompt treatment. A proper medical infrastructure must be set up in case of an outbreak because a substantial portion of the people in our nation lives in rural areas. As one of the most frequent epidemics in our nation and the cause of a high fatality rate, dengue needs to be predicted well in advance so that medical resources may be set up on time and the fatality rate is lowered. Although dengue does not require a particular course of therapy or a lengthy course of treatment, it is crucial to identify and address the condition quickly.

The technology seeks to forecast the outbreak of an epidemic (dengue), allowing the health sector to prepare the necessary resources in advance. Recurrent neural networks will be used by the system to produce predictions. Climate, pollution, and the numbers of people who have been diagnosed with dengue in prior years are only a few of the variables included in the forecast process [43]. The information was gathered from a number of official sources as well as medical facilities. In the future, if comparable environmental circumstances exist, the model will learn from this data and forecast the likelihood of an outbreak. The outcome of the potential outbreak will be displayed by using Google heatmaps to indicate the potential dengue-endemic areas.

In order to implement more effective vector management techniques, it is critically necessary to develop mechanisms that can more accurately predict dengue cases. A unique hybrid architecture that combines the advantages of recurrent neural network and convolutional neural networks is used to identify the weekly dengue incidence [44]. In dengue predicting method, this hybrid architecture beats the other frequently used deep learning models. The suggested models are also compared to cutting-edge studies in the literature, illuminating how significantly dengue forecasting can benefit from our hybrid models' use of recurrent networks and convolutional layers.

2.7 Ensemble techniques

In machine learning techniques, an ensemble of basic models is combined to provide an ideal prediction model and/or learner that enhances the outcome. In Bagging the training data is divided into multiple copies, and after that, each copy is used to train a different classifier. By merging the predictions made by various classifiers, the final prediction is formed. Homogenous weak classifiers are constructed in series to correct the classifier errors [45]. The first model was developed using the whole dataset, while the second model used the predictions from the first model. All of the weak classifiers are combined to create the final classifier. Staggering process builds heterogeneous weak classifiers and the stacking ensemble approach works using a meta model. This meta model develops the ability to mix base model predictions. The level 0 classifier output is used as training data for the level 1 classifier, which will then try to approach the same goal function, according to the stacking combining technique. An ensemble learning technique called a random forest consists of numerous decision trees, or different classifiers [46]. Both classification and regression issues can be resolved using it. The following steps are used to make a random forest: Collect the required input data, then for each tree select the randomised subsets of the data as candidate inputs, now the candidate predictors are used to build a tree, final step is to identify the prediction from each tree in the forest and calculate the average of all predictions. Both numerical and categorical data can be handled by a random forest, but the method performs best with continuous response variables. It is broadly used with logistic regression, support vector machines and linear regression.

A research study in south China was conducted based on statistically approach to predict dengue outbreak in real time [47]. The main idea behind this research was to developed an ensemble forecast model for predicting dengue outbreak in real time. The prediction model has used dengue data, climate data, search query data and social media data as an input data in the prediction model. This model undergoes four phases namely Raw data, Process data, Ensemble data and Aggregated data. The outcome of this survey paper is that EPRA prediction model shows best performance for 1-2 week ahead. The model has achieved 94% accuracy. This prediction model can be further implemented for central region to other regions of China.

Iqbal et al. [48] concentrated on building a model that can effectively recognise the extent of the dengue outbreak in a large number of specimens at once. Seven well-known machine learning systems were evaluated for their capacity to predict dengue outbreaks. Eight different performance parameters are used to assess these approaches' effectiveness. The best classification accuracy was reported to be 92% using the LogitBoost ensemble model, with specificity and sensitivity of 94% and 90%, respectively.

The study that shows how environmental elements including temperature, humidity, precipitation, and others affect dengue transmission. Transmission of the dengue virus is more likely in regions with higher vapour pressure and precipitation rates. It made use of classification techniques to identify the key traits that caused dengue to spread. Ensemble learning technique was employed in hybrid integrations to identify the characteristics that spread the illness caused because of dengue [49].

The reliability and accuracy of a number of automated Aedes larvae detection techniques that have been previously proposed were insufficient [50]. Therefore, an automated approach was proposed which are very effective and accurate in distinguishing the Aedes larvae by providing an accuracy of about 99%. When focus on accuracy, this methodology is far better when compared to any of the earlier techniques.

3. ARTICLE SELECTION PROCESS

Indian Healthcare system is broadly operating in two forms namely public as well as private sector. Indian people are categorised as rich class, middle class, lower middle class and below poverty line. The people of India are living in urban, semi urban, rural and remote areas. The wealthy people afford private healthcare facilities easily whereas lower middle and poor class avail public healthcare facilities at the time of their health problems. Further, in rural and remote areas where both private and public healthcare facilities are seldom available, an artificial intelligence-based healthcare facility can be beneficial for the residents. Also, the rural population in India is the most affected by vector borne diseases.

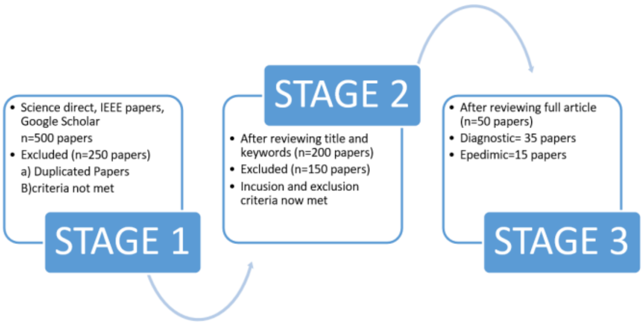
• Science direct, IEEE papers, Google Scholar
n=500 papers
• Excluded (n=250 papers)
a) Duplicated Papers
B)criteria not met
STAGE 1
STAGE 2
• After reviewing title and keywords (n=200 papers)
• Excluded (n=150 papers)
• Incusion and exclusion criteria now met
• After reviewing full article (n=50 papers)
• Diagnostic= 35 papers
• Epedimic=15 papers
STAGE 3

Figure 2. Stages of paper selection

Dengue is such a vector borne disease which spreads in some villages every year. The prevention and control of this disease is still a challenging task for the government as well as for the healthcare agencies. The goal of the research is to prepare machine learning based prediction model for early detection of dengue disease. For the welfare of humankind, a prediction model was proposed which can diagnose dengue at the early stages. To develop a robust and efficient model a thorough literature review is required that can facilitate the model building.

For this survey, searches are performed on online databases, with the help of Scopus, PubMed, Science Direct and Google Scholar. Figure 2 shows the various stages of paper selection. We applied the subsequent inclusion criteria: keyword phrases include infectious infections, red patches, severe headaches, night sweats, and dengue. Journal articles, conference papers, books, and book chapters were all included in the document type, and deep learning, AI, and machine learning were all included in the search for theme areas. Not a part of disease diagnosis, absence of English language, absence of conclusive, and not being in the field of ML, deep learning, or AI are all exclusions. 500 results came up in the initial search. 50 publications were chosen for our study once inclusion criteria were applied.

4. RESULTS AND DISCUSSION

The dataset for dengue prediction highly relies on labelled data and hence the supervised Machine algorithms are chosen for the study. Before deciding which algorithm or model is ideal for a certain task, it is often suggested to test a few and evaluate their performance. Prior to using any machine learning or deep learning method, it is crucial to correctly pre-process and normalise the data. The Table 1 lists few pros and cons for machine learning techniques depending on the specific dataset and problem being solved depending on which the model selection can be done. The Table 2 describes the common features of dengue defined as binary data type.

Survey articles were taken on the basis of the above analysation in predicting dengue for diagnostic models for infectious diseases. Out of the 41 survey articles, it was discovered that 20% of research publications used SVM, 7% used ANN, 7% used KNN and 5% used LR techniques. The remaining 15% DT, 20% NB, "boosting" ensemble approach 12% and 5% CNN approaches. Other machine learning methods include deep learning, J48, adaptive network-based fuzzy inference systems (ANFIS), and random forest. (VQ). Figure 3 shows the usage of various ML techniques that was surveyed.

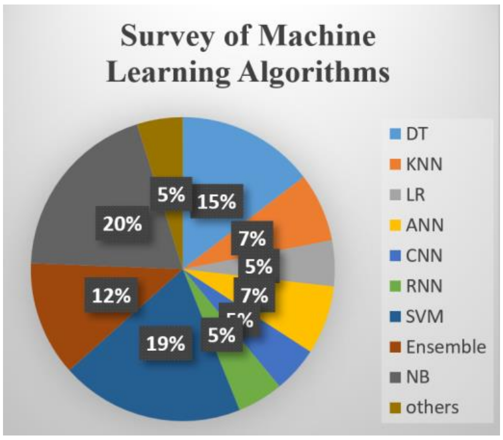

Figure 3. Frequency of ML algorithms used in the study

The machine learning algorithms discussed in the study are evaluated to identify the performance of the model and also to test the generalization capabilities using Cross validation and K-Fold validation. The Cross-validation method is a statistical approach for comparing and evaluating algorithms by splitting the data into two sets where, one set of data is used for training and the other set is used for testing. The K-Fold validation on the other hand, splits the data into k number of folds which is used to evaluate the model with new data set. Precision and F1 score are the performance metrics used to evaluate the algorithm's effectiveness. The Table 3 shows the evaluation of ML techniques for Cross validation, K-Fold Validation, Precision and F1 score. Table 2 parameters are input to the algorithm. The values are given in binary. The output of each algorithm is shown in Table 3.

The Figure 4 shows the comparative analysis of dengue disease prediction models of various ML algorithms like logistic regression, K-nearest neighbour, Support Vector machine, Decision tree, Naïve Bayes and Artificial Neural Network.

Table 1. Comparisons of ML algorithms

Algorithm	Pros	Cons
Logistic Regression	Simple and widely used algorithm for linear data.	Can be prone to overfitting if dataset is too complex.
K-Nearest Neighbours	Non-parametric method that works well with complex decision boundaries. Can handle both numerical and categorical data.	Sensitive to choose k. Can be computationally expensive for large datasets.
Support Vector Machines	Powerful algorithm that can handle both linear and non-linear datasets. Can work well for datasets with a clear margin of separation between classes.	Can be computationally expensive. May not perform well on datasets with noisy or overlapping classes.
Decision Trees	Simple and interpretable. Can handle both numerical and categorical data.	Can be prone to overfitting. May not perform well on datasets with imbalanced classes.
Naive Bayes	Simple and computationally efficient. Can handle high dimensional data. Assumes predictor variables are independent.	Assumes predictor variables are independent, which may not be true in all cases.
Neural Networks	Can handle non-linear data and learn complex patterns. Can learn from large amounts of data.	Can be prone to overfitting and can be computationally expensive.
Ensemble Techniques	Can work well with noisy or imbalanced classes.	Can be computationally expensive.
ANN	Can handle non-linear data and learn complex patterns.	Can be prone to overfitting and can be computationally expensive.
CNN	Can learn complex features from raw input data. Works well for datasets with spatial or temporal dependencies.	May require large amounts of data to learn patterns.
LSTM	Can learn from sequences of data and handle long-term dependencies. Can work well for time-series prediction.	Can be computationally expensive.

Table 2. Common features in dengue fever

Feature	Description
Logistic Regression	Simple and widely used algorithm for linear data.
Fever	Whether the patient is experiencing fever or not.
Headache	Whether the patient is experiencing headache or not.
Muscle Pain	Whether the patient is experiencing muscle pain or not.
Joint Pain	Whether the patient is experiencing joint pain or not.
Rash	Whether the patient is experiencing rash or not.
Nausea	Whether the patient is experiencing nausea or not.
Vomiting	Whether the patient is experiencing vomiting or not.
Fatigue	Whether the patient is experiencing fatigue or not.
Eye Pain	Whether the patient is experiencing eye pain or not.
Body Aches	Whether the patient is experiencing body aches or not.
Diarrhoea	Whether the patient is experiencing diarrhoea or not.
Loss of Appetite	Whether the patient is experiencing loss of appetite or not.
Bleeding	Whether the patient is experiencing bleeding or not.
Low White Blood Cell Count	Whether the patient has low white blood cell count or not.
Low Platelet Count	Whether the patient has low platelet count or not.
Positive Dengue NS1 Antigen Test	Whether the patient has tested positive for dengue NS1 antigen or not.

Table 3. Comparative analysis of the dengue prediction model of various ML algorithms

Algorithm	Cross Validation	K-Fold Validation	Precision	F1
LR	96.5%	96.5%	97.0%	96.0%
KNN	97.2%	96.5%	97.0%	97.%
SVM	98.5 %	97.5%	98.2%	98.0%
DT	97.4 %	97.1 %	96.0%	96.0%
NB	97.3 %	97.2%	99.1%	99.2%
ANN	97.5%	97.3%	98.2%	97.4%

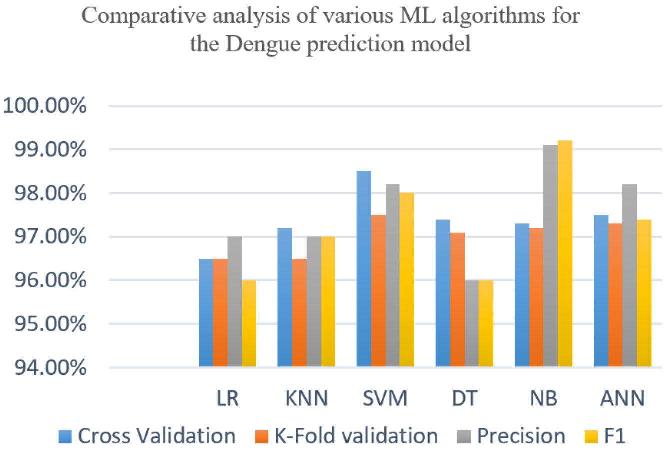

Figure 4. Comparative analysis of various ML algorithms for the dengue prediction model

The study is carried out with the same binary text input dataset for all algorithms mentioned in Table 3. Among which Naive Bayes can outperform other algorithms especially when dealing with text data. As a result, Naive bayes algorithm provides output with 97.3% through Cross Validation and 97.2% through K-Fold Validation. The Precision value is 99.1% and F1 score is 99.2% from the above analysis. SVM also provides overall performance with cross validation of 98.5%, K-Fold validation of 97.5%, Precision of 98.2% and F1 Score of 98.0%.

5. CONCLUSIONS

Infectious disease diagnosis takes time and money. Due to issues like population density, poverty and a lack of access to healthcare the developing countries frequently have a higher burden of infectious diseases. Hence Machine Learning offers a number of alternatives that can give precise results more quickly with a better diagnosis compared human labour. There are more recent developments in ML algorithms are thus examined and concentrated on how they might be used to detect the infectious diseases. These algorithms have the potential to revolutionize how diseases are detected, diagnosed and treated, leading to more accurate and timely interventions. From the study, Supervised ML algorithms are frequently employed for diagnosis that predicts the disease with more accuracy. Multiclass and multilabel classification issues can be successfully handled with Naive Bayes handling tasks like text categorization when there are several classes or where each instance can have multiple labels attached to it. SVMs are frequently employed in a wide range of machine learning applications and have proven successful in many instances. From the findings, SVM models enhance the predictive model's overall performance on the basis of validation, precision and F1 score. In conclusion, the integration of machine learning in disease diagnosis would lead to a collaborative healthcare landscape, where human expertise and AI capabilities complement each other for improved patient care, faster diagnoses, and better treatment outcomes.

REFERENCES

[1] Guzman, M.G., Jaenisch, T., Gaczkowski, R., et al. (2010). Multi-country evaluation of the sensitivity and specificity of two commercially-available NS1 ELISA assays for dengue diagnosis. PLoS Neglected Tropical Diseases, 4(8): e811. https://doi.org/10.1371/journal.pntd.0000811

[2] Dey, S.K., Rahman, M.M., Howlader, A., Siddiqi, U.R., Uddin, K.M.M., Borhan, R., Rahman, E.U. (2022). Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach. PLoS One, 17(7): e0270933. https://doi.org/10.1371/journal.pone.0270933

[3] Meyfroidt, G., Güiza, F., Ramon, J., Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. Best Practice & Research Clinical Anaesthesiology, 23(1): 127-143. https://doi.org/10.1016/j.bpa.2008.09.003

[4] Dahiwade, D., Patle, G., Meshram, E. (2019). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. https://doi.org/10.1109/ICCMC.2019.8819782

[5] Kim, E., Rubinstein, S.M., Nead, K.T., Wojcieszynski, A.P., Gabriel, P.E., Warner, J.L. (2019). The evolving use of electronic health records (EHR) for research. In Seminars in Radiation Oncology, 29(4): 354-361. https://doi.org/10.1016/j.semradonc.2019.05.010

[6] Asan, O., Bayrak, A.E., Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. Journal of Medical Internet Research, 22(6): e15154. https://doi.org/10.2196/15154

[7] Pandey, S.K., Janghel, R.R. (2019). Recent deep learning techniques, challenges and its applications for medical healthcare system: A review. Neural Processing Letters, 50: 1907-1935. https://doi.org/10.1007/s11063-018-09976-2

[8] Cruz, J.A., Wishart, D.S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2: 117693510600200030. https://doi.org/10.1177/117693510600200030

[9] Chandru, A.S., Seetharam, K. (2020). Framework for efficient transformation for complex medical data for improving analytical capability. International Journal of Electrical and Computer Engineering, 10(5): 4853.

[10] Caicedo-Torres, W., Montes-Grajales, D., Miranda-Castro, W., Fennix-Agudelo, M., Agudelo-Herrera, N. (2017). Kernel-based machine learning models for the prediction of dengue and chikungunya morbidity in Colombia. In Advances in Computing: 12th Colombian Conference, CCC 2017, Cali, Colombia, pp. 472-484. https://doi.org/10.1007/978-3-319-66562-7_34

[11] Mussumeci, E., Coelho, F.C. (2020). Machine-learning forecasting for Dengue epidemics-comparing LSTM, Random Forest and Lasso regression. Medrxiv, 2020-01. https://doi.org/10.1101/2020.01.23.20018556v1

[12] Kramer, O. (2013). Dimensionality reduction with unsupervised nearest neighbors. Intelligent Systems Reference Library, 51: 13-23. https://doi.org/10.1007/978-3-642-38652-7

[13] Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3): 175-185. https://doi.org/10.1080/00031305.1992.10475879

[14] Vijayakumar, V., Malathi, D., Subramaniyaswamy, V., Saravanan, P., Logesh, R. (2019). Fog computing-based intelligent healthcare system for the detection and prevention of mosquito-borne diseases. Computers in Human Behavior, 100: 275-285. https://doi.org/10.1016/j.chb.2018.12.009

[15] Pisner, D.A., Schnyer, D.M. (2020). Support vector machine. In Machine learning, pp. 101-121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[16] Noble, W.S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12): 1565-1567. https://doi.org/10.1038/nbt1206-1565

[17] Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5): 3.

[18] Auria, L., Moro, R.A. (2008). Support vector machines (SVM) as a technique for solvency analysis. https://ideas.repec.org/p/diw/diwwpp/dp811.html#:~:text=This%20paper%20introduces%20a%20statistical%20technique%2C%20Support%20Vector,accuracy%20of%20company%20classification%20into%20solvent%20and%20insolvent.

[19] Nordin, N.I., Sobri, N.M., Ismail, N.A., Zulkifli, S.N., Abd Razak, N.F., Mahmud, M. (2020). The classification performance using support vector machine for endemic dengue cases. In Journal of Physics: Conference Series, 1496(1): 012006. https://doi.org/10.1088/1742-6596/1496/1/012006

[20] Kesorn, K., Ongruk, P., Chompoosri, J., Phumee, A., Thavara, U., Tawatsin, A., Siriyasatien, P. (2015). Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. PloS One, 10(5): e0125049. https://doi.org/10.1371/journal.pone.0125049

[21] Fathima, A., Manimegalai, D. (2012). Predictive analysis for the arbovirus-dengue using svm classification. International Journal of Engineering and Technology, 2(3): 521-527.

[22] Dourjoy, S.M.K., Rafi, A.M.G.R., Tumpa, Z.N., Saifuzzaman, M. (2021). A comparative study on prediction of dengue fever using machine learning algorithm. In Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020, pp. 501-510. https://doi.org/10.1007/978-981-15-4218-3_49

[23] Murthy, S.K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery, 2: 345-389. https://doi.org/10.1023/A:1009744630224

[24] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6): 275-285. https://doi.org/10.1002/cem.873

[25] Dudkina, T., Meniailov, I., Bazilevych, K., Krivtsov, S., Tkachenko, A. (2021). Classification and prediction of diabetes disease using decision tree method. In IT&AS, pp. 163-172.

[26] Kalansuriya, C.S., Aponso, A.C., Basukoski, A. (2020). Machine learning-based approaches for location based dengue prediction. In Fourth International Congress on Information and Communication Technology: ICICT 2019, London, 1: 343-352. https://doi.org/10.1007/978-981-15-0637-6_29

[27] Pravin, A., Jacob, T.P., Nagarajan, G. (2020). An intelligent and secure healthcare framework for the prediction and prevention of Dengue virus outbreak using fog computing. Health and Technology, 10: 303-311. https://doi.org/10.1007/s12553-019-00308-5

[28] Rao, V.S.H., Kumar, M.N. (2011). A new intelligence-based approach for computer-aided diagnosis of dengue fever. IEEE Transactions on Information Technology in Biomedicine, 16(1): 112-118. https://doi.org/10.1109/TITB.2011.2171978

[29] Sammut, C., Webb, G.I. (Eds.). (2011). Encyclopedia of Machine Learning. Springer Science & Business Media.

[30] Dada, E.G., Bassi, J.S., Chiroma, H., Adetunmbi, A.O., Ajibuwa, O.E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. Heliyon, 5(6). https://doi.org/10.1016/j.heliyon.2019.e01802

[31] Zhang, W., Gao, F. (2011). An improvement to naive bayes for text classification. Procedia Engineering, 15: 2160-2164. https://doi.org/10.1016/j.proeng.2011.08.404

[32] Harumy, T.H.F., Chan, H.Y., Sodhy, G.C. (2020). Prediction for dengue fever in Indonesia using neural network and regression method. In Journal of Physics: Conference Series, 1566(1): 012019. https://doi.org/10.1088/1742-6596/1566/1/012019

[33] Peng, L.H., Zhou, L.Q., Chen, X., Piao, X. (2020). A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. Frontiers in Bioengineering and Biotechnology, 8: 40. https://doi.org/10.3389/fbioe.2020.00040

[34] Azam, N., Salim, M., Yap, B.W., et al. (2021). Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Scientific Reports, 11(1): 939. https://doi.org/10.1038/s41598-020-79193-2

[35] Somwanshi, H., Ganjewar, P. (2018). Real-time dengue prediction using naive bayes predicator in the IoT. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 725-728. https://doi.org/10.1109/ICIRCA.2018.8596796

[36] Caicedo-Torres, W., Paternina, Á., Pinzón, H. (2016). Machine learning models for early dengue severity prediction. In Advances in Artificial Intelligence-IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25, 2016, Proceedings 15, pp. 247-258. https://doi.org/10.1007/978-3-319-47955-2_21

[37] Zhang, Z., Zhang, K., Khelifi, A. (2018). Multivariate time series analysis in climate and environmental research, p. 287. https://doi.org/10.1007/978-3-319-67340-0

[38] Basheer, I.A., Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. Journal of Microbiological Methods, 43(1): 3-31. https://doi.org/10.1016/S0167-7012(00)00201-3

[39] Liu, T., Fang, S., Zhao, Y., Wang, P., Zhang, J. (2015). Implementation of training convolutional neural networks. arXiv preprint arXiv:1506.01195. https://arxiv.org/abs/1506.01195

[40] Mishra, S., Kumar, R., Tiwari, S.K., Ranjan, P. (2022). Machine learning approaches in the diagnosis of infectious diseases: A review. Bulletin of Electrical Engineering and Informatics, 11(6): 3509-3520. https://doi.org/10.11591/ijeecs.v26.i1.pp352-361

[41] Musha, A., Al Mamun, A., Tahabilder, A., Hossen, M.J., Jahan, B., Ranjbari, S. (2022). A deep learning approach for COVID-19 and pneumonia detection from chest X-ray images. International Journal of Electrical & Computer Engineering (2088-8708), 12(4).

[42] Gambhir, S., Malik, S.K., Kumar, Y. (2018). The diagnosis of dengue disease: An evaluation of three machine learning approaches. International Journal of Healthcare Information Systems and Informatics (IJHISI), 13(3): 1-19. https://doi.org/10.4018/IJHISI.2018070101

[43] Chovatiya, M., Dhameliya, A., Deokar, J., Gonsalves, J., Mathur, A. (2019). Prediction of dengue using recurrent neural network. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 926-929. https://doi.org/10.1109/ICOEI.2019.8862581

[44] Zhao, X., Li, K., Ang, C.K.E., Cheong, K.H. (2023). A deep learning based hybrid architecture for weekly dengue incidences forecasting. Chaos, Solitons & Fractals, 168: 113170. https://doi.org/10.1016/j.chaos.2023.113170

[45] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14: 241-258. https://doi.org/10.1007/s11704-019-8208-z

[46] Petkovic, D., Altman, R., Wong, M., Vigil, A. (2018). Improving the explainability of Random Forest classifier–user centered approach. In Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium, pp. 204-215. https://doi.org/10.1142/9789813235533_0019

[47] Guo, P., Zhang, Q., Chen, Y., Xiao, J., He, J., Zhang, Y., Wang, L., Liu, T., Ma, W. (2019). An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. Science of The Total Environment, 647: 752-762. https://doi.org/10.1016/j.scitotenv.2018.08.044

[48] Iqbal, N., Islam, M. (2019). Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica, 43(3). https://doi.org/10.31449/inf.v43i3.1548

[49] Gangula, R., Thirupathi, L., Parupati, R., Sreeveda, K., Gattoju, S. (2023). Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns. Materials Today: Proceedings, 80: 3458-3463. https://doi.org/10.1016/j.matpr.2021.07.270

[50] Hossain, M.S., Raihan, M.E., Hossain, M.S., Syeed, M.M., Rashid, H., Reza, M.S. (2022). Aedes larva detection using ensemble learning to prevent dengue endemic. BioMedInformatics, 2(3): 405-423. https://doi.org/10.3390/biomedinformatics2030026