Journal of New Materials for Electrochemical Systems
Vol. 28, No. 1, January 2025, pp 84-96
Journal homepage: https://www.newmaterials.ca/

Biosensor based Detection of Early Spread of Vector Borne Disease with Personalized Treatment strategies using Machine Learning

Suresh Maruthai[1], Balamurugan K S[2], Tamilvizhi Thanarajan[3], Surendran Rajendran[4,*]

[1] Department of Electronics and Communication Engineering, St. Joseph's College of Engineering, Chennai-600 119, India. maruthaihsuresh@gmail.com
[2] Department of ECE, Karpaga Vinayaga College of Engineering and Technology, Chengalpattu, Tamil Nadu, India. profksbala@gmail.com
[3] Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, 600123, India.
[4] Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, 602105, India. dr.surendran.cse@gmail.com

Corresponding Author Email: dr.surendran.cse@gmail.com

https://doi.org/10.14447/jnmes.v28i1.a09

Received: March 19, 2024
Accepted: Sep 09, 2024

Keywords:
Biosensor, vector borne diseases, Support vector machine, Random Forest, Naïve Bayes

ABSTRACT

Vector-borne diseases in India show growing patterns which strongly affect the population of the nation. The government faces a major obstacle in disease prevention efforts. Every year many people throughout India suffer from these illnesses. The physical differences between geographic regions and ways of life make it difficult for current strategies to control diseases during their initial development stages. The project focuses on creating advanced methods based on machine learning to diagnose diseases caused by vectors. The planned investigation targets dengue rather than other vector-borne diseases because it has emerged as one of the most dominant pathologies in contemporary years. A total of five stages make up the proposed methodology beginning with Data Transformation after which Preprocessing occurs followed by Feature Scaling and Normalization and finally Dataset Partitioning to allow Model Development and Prediction. The proposed model brings forth an ability to identify dengue fever development throughout its stages. The proposed solution stands out because it identifies dengue fever during early stages while determining the disease severity using patient clinical information. A test of the model used Support Vector Machine (SVM), Decision Tree and Gaussian Naïve Bayes Classifier, Logistic Regression and Random Forest Classifier algorithms for evaluation. A biosensor was used for extensive testing and validation which enabled the suggested technique to produce a 97.5% accuracy rate. The Gaussian Naïve Bayes classifier achieved 97.5% accuracy although it had a root mean square error value of zero.

1. INTRODUCTION

Human beings are considered the pinnacle of God's creation, constantly striving to improve their lifestyles through advancements in science and technology, which have made life easier. Despite these advancements, challenges persist in protecting human life from evolving diseases transmitted through humans, animals, and environmental factors. Medical science has developed diagnostic systems to identify and treat diseases, playing a crucial role in saving lives. With the advent of AI and ML, intelligent machines now monitor health parameters and patient status continuously. The field of vector-borne disease research, however, remains in its infancy, requiring extensive studies on vectors, environmental factors, and the development of early detection and prevention technologies to prevent outbreaks. [1,2].

Vector-borne diseases, transmitted through bites of infected vectors like mosquitoes, ticks, and fleas, pose a significant threat to global public health. The complex interplay of environmental factors, vector dynamics, and human behavior makes early detection and effective treatment challenging. As these diseases spread, innovative approaches are needed to identify outbreaks early and tailor treatment strategies. This research harnesses machine learning to detect early disease growth and design personalized treatments. Traditional surveillance methods struggle with these diseases' dynamic nature, while ML can analyze large, diverse datasets, improving early intervention and patient outcomes. The focus is on developing robust predictive models and tailoring treatment plans based on individual profiles. [3,4]. By integrating complex data sources and using advanced ML algorithms, we aim to create a framework for early outbreak detection and personalized treatment recommendations for vector-borne diseases. This research could revolutionize

treatment approaches, enhancing therapeutic outcomes and minimizing adverse effects. Real-time data from IoT devices and wearable technologies will enable continuous monitoring, providing a proactive, dynamic approach to public health management.

This research aims to advance precision medicine for vector-borne diseases and strengthen healthcare resilience. Our research presents opportunities for developing ahead-of-time public health strategies based on machine learning methods to defend vulnerable groups against growing vector-borne diseases. [5] Human hosts and vectors together with pathogens create complex disease transmissions that result in major health system strains across the globe. Malaria together with dengue and Zika and Lyme borreliosis diseases cover wide geographic areas while displaying varying transmission patterns. Practices of traditional disease detection together with treatment methods create delays that cause higher illness severity leading to more deaths. Machine learning (ML) technologies achieved new levels of data analysis speed and accuracy through their advent. The research goals utilize machine learning technology to achieve vector-borne disease early diagnosis and customized therapeutic approaches which combat persistent public health crises. [6]

Our goal is to merge different datasets that contain environmental elements together with patient health information in order to develop a complete predictive model for disease outbreaks and intervention customization. The immediate need to conduct this research emerges because vector-borne diseases intensify due to climate change alongside urbanization and increasing international mobility. The current disease control techniques prove inadequate so healthcare needs to transition toward personalized disease prevention approaches. Our first objective involves building predictive models through ML algorithms to analyze patterns among historical data.[7]

The purpose of these models is to detect early warning signs of disease outbreak by studying vector populations with environmental factors together with human behavior elements which allows public health interventions to occur promptly and with precise targeting. The method of adopting a proactive approach proves essential to stop transmission patterns and limit major disease outbreaks. Treatment strategies should be individualized through research that considers genetics as well as lifestyle and medical background components for every patient. Through this therapy method clinical outcomes achieve peak performance while reducing adverse reactions to improve patient health. To enhance real-time access, we integrate data from IoT devices and wearable technologies for continuous monitoring and early outbreak identification, facilitating timely interventions.

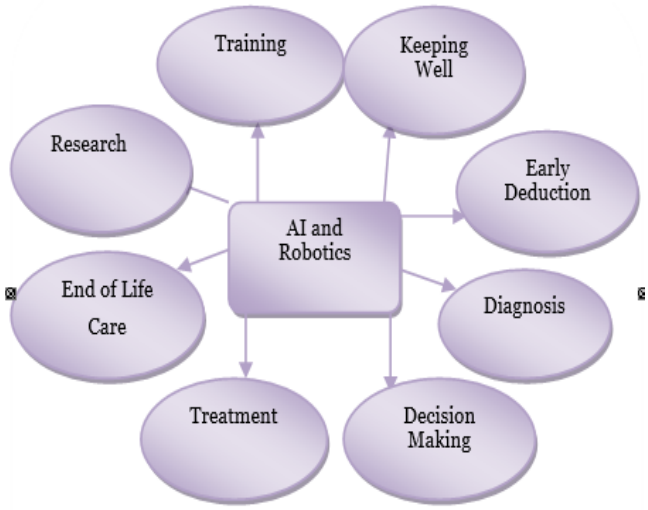

Figure 1. Artificial Intelligence and Robotics

This research harnesses machine learning to revolutionize the detection and treatment of vector-borne diseases.[8,9]

By combining predictive modeling with personalized treatment strategies, we aim to develop a resilient healthcare system to address evolving challenges in the 21st century.[10]

1.1 Application of machine learning in healthcare

Healthcare has become a major application area for machine learning, driven by advances in technology, new medical devices, and drugs. This sector not only saves lives but also provides employment to millions globally. Information technology plays a significant role, with smart, internet-connected medical devices monitoring multiple patients and alerting medical teams in real-time. Machine learning is increasingly accepted in healthcare, becoming a

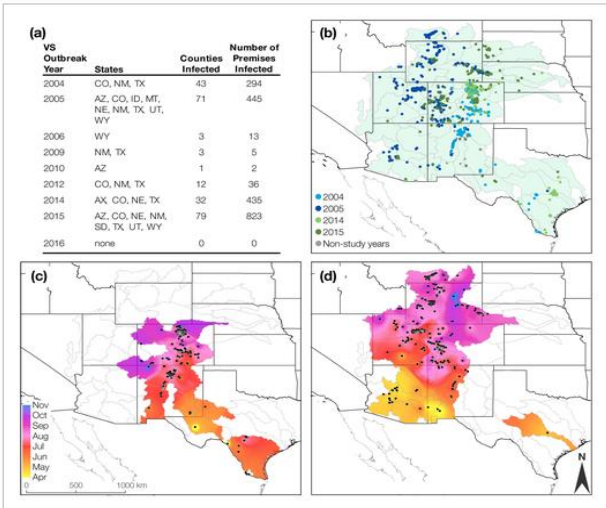

Figure 2. The illustration presents various aspects:

All reported cases from 2004 to 2015 are thoroughly displayed under four categories which include annual breakdown alongside state and county data and the types of constructed facilities with reported infections.

vital tool for healthcare staff and doctors.[11,12] It analyzes vast amounts of data on diseases, offering valuable suggestions for treatment represented in Figure 2.

Blue color marks the time period between 2004 and 2005 and the green color indicates the years from 2014 to 2015. The illustration gives special attention to infection cases that occurred during 2004. The depiction focuses mostly on 2005 events to show how disease trends spread from south to north through highlighted areas with distinct gauge markings.

Table 1 Comparison of Related work

Author	Title	Technologies	Year
Shubham Kumar, et al, [15]	"Modeling climate change impacts on vector-borne disease using machine learning models: Case study of Visceral leishma"	RBF-kernel-based-SVR model	2024
Daozhou Gao, et al, [16]	"Vector-borne disease models with Lagrangian approach"	develop a multi-group and multi-patch model	2024
Ritesh Chandra, et al, [17]	"Semantic web-based diagnosis and treatment of vector-borne diseases using SWRL rules"	Semantic Web Rule Language (SWRL) rules	2023
Inderpreet Kaur, et al, [18]	"Artificial Intelligence Techniques for Predictive Modeling of Vector-Borne Diseases and its Pathogens: A Systematic Review"	Machine Learning & Deep Learning techniques	2022

After exploring several established techniques for classifying vector-borne diseases, identified significant gaps in innovation. To address these challenges, I have proposed a novel method titled "Detection of Early Spread of Vector Borne Disease with Personalized Treatment Strategies Using Machine Learning." This approach aims to revolutionize disease classification by leveraging advanced machine learning techniques for early detection and tailored treatment strategies.Top of FormBottom of Form

2. EXPERIMENTAL SECTION

The critical barrier to vector-borne disease prediction emerges from multiple variables which affect disease spread patterns. Disease transmission is strongly influenced by environmental elements together with socioeconomic conditions and human conduct patterns and these components exhibit major variations across regional areas and within specific time frames. Inaccurate predictions stem from incomplete high-quality data collections which generate unreliable mathematical predictions due to inconsistent sample information. The development of accurate models for forecasting does not remain feasible across time due to evolving pathogen strains and modification in ecosystems. The solution to these problems needs multidimensional coordination between machine learning technologies and reliable data gathering and examination methods.

Development of machine learning algorithms represents our solution to address the major forecasting hurdles of vector-borne diseases. Advanced machine learning methods will serve as our approach for analyzing factors which influence disease transmission because of their analytical capability with complex and variable elements. We build precise and dependable predictive models through analysis of extensive information that combines environment data with social economic indicators and behavioral patterns of humans. Our approach includes specific techniques which deal with data inconsistencies and vacant areas to maintain the reliability of our forecasting predictions. This solution will boost outbreak predictions which will lead to enhanced intervention strategies that decrease the impact vector-borne diseases have on affected populations.

2.1 Information Acquisition:

The program requires a complete set of patient information that contains demographics coupled with medical histories and clinical vector-borne disease data. The research team must gather information about environmental elements that include temperature measurements alongside humidity data and precipitation rates and land use patterns since these variables determine vector living spaces and disease transmission rates.

2.2 Vector Population Dynamics:

Investigate vector populations together with their breeding areas and geographical distribution.

2.3 Data preprocessing:

The collected data requires cleaning followed by data preprocessing to handle missing values and outliers before it becomes suitable for machine learning purposes. Normalize numeric values while standardizing their characters and use proper coding methods for categorical variables.

2.4 Selection of functions:

Feature selection methods should identify which variables most impact disease transmission alongside patient response to treatment.

2.5 Machine learning models:

The development of predictive models for early detection by using Random Forests and Support Vector Machines and Neural Networks procedures will be performed. The modeling process should use historical data for training together with validation subsets for performance optimization. Multi-model systems should be used to enhance the reliability of predictive outcomes.

2.6 Personalized treatment model:

The team will build a system that generates personalized medical solutions based on particular patient information. Include genetic data and life-style elements and historical medical records when developing the treatment recommendation model. Machine learning algorithms with data-explaining features help healthcare specialists understand their decision algorithms.

2.7 Real-time monitoring integration:

A system should collect real-time data streams that originate from IoT devices and wearables to perform continuous patient tracking. Machine learning models will receive information through implemented data pipelines that connect monitoring devices to the models.

2.8 Verification and evaluation:

The assessment of predictive models depends on retrospective data analysis where forecasted outbreaks are compared to historical records. Measure how well a customized treatment method performs using medical results and patient_responses to determine its precision and response rate and specificity.

2.9 Prospective study:

Carry out future examinations in regions with high vector-borne diseases to verify how quickly the proposedframework performs in practice.

2.10 Ethical considerations:

The process should comply with ethical standards for both data collection procedures and database management and patient privacy rules. The researchers secure required Institutional Review Board (IRB) approvals and obtain participant-informed consent for every individual included in the study.

2.11 Analysis and interpretation:

Assess experimental outcomes based on the combination of timely detection methods and individualized treatment methods. Manufacture a practical analysis of model results which healthcare professionals can use in Figure 3.

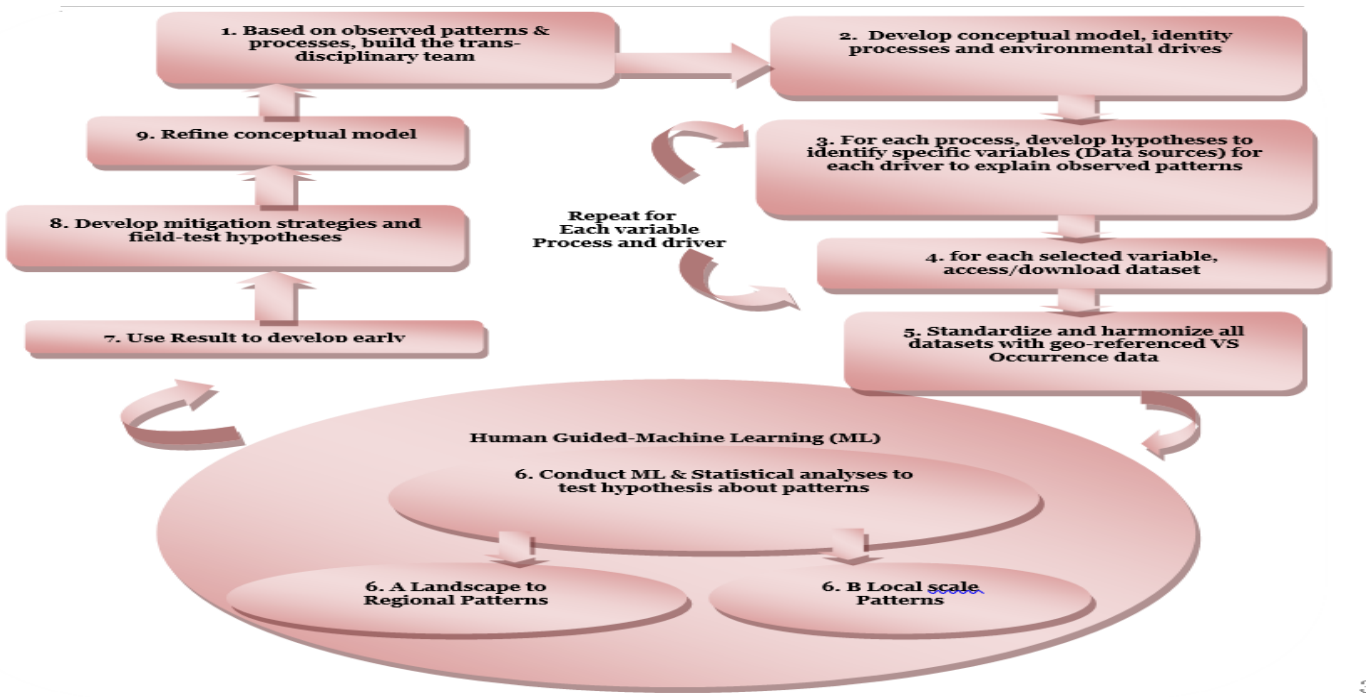

Figure 3. Workflow approach to predictive disease

The team behind the study comprises disease specialists together with environmental scientists and experts in environmental informatics and machine learning experts and professionals who specialize in big data analytics (Fig. 3). Our method works to analyze how patterns and processes connect with individuals starting from the initial period up to sickness development. Research indicates that the exact time and space measurement of a certain geographic region results from different factors working together with a vector-based system which triggers infectious processes. The framework combines big data from the ecosystem and integrates knowledge through expert information and machine learning methods. Three distinct sections appear within the diagram: blue boxes feature disease specialist and ecological contributions and pink boxes display ecologic information expertise while yellow boxes dedicate themselves to human-driven machine learning approaches.

The analysis began with local and spatial processes such as vertical and horizontal transmission and insect overwintering behavior and animal-to-animal transmission and insect-to-animal transmission. Subsequently six environmental factors were identified for their potential impact on disease ecology systems based on research findings. The VS system utilized its expertise to choose variables from each driving factor that would affect disease process patterns or responses during the emergence of VS occurrence patterns. The method enables investigators to explore many potential variables before focusing on crucial biological variables which matter in analysis.

The analysis began with local and spatial processes such as vertical and horizontal transmission and insect overwintering behavior and animal-to-animal transmission and insect-to-animal transmission. Subsequently six environmental factors were identified for their potential impact on disease ecology systems based on research findings. The VS system utilized its expertise to choose variables from each driving factor that would affect disease process patterns or responses during the emergence of VS occurrence patterns. The method enables investigators to explore many potential variables before focusing on crucial biological variables which matter in analysis [13].

The proposed research framework divides its work into four distinct parts starting from Data Collection and proceeding through Data Preprocessing and Data Splitting and Classification before Evaluation with a Confusion Matrix. In the starting stage of data collection, information is gathered From Kaggle. The dataset compiled for predicting vector-borne diseases encompasses detailed information on symptoms, prognosis, and diagnostic parameters for eleven specific diseases. Each disease, such as Chikungunya, Dengue, Zika, Yellow Fever, Rift Valley Fever, West Nile Fever, Malaria, Tungiasis, Japanese Encephalitis, Plague, and Lyme Disease, is characterized by distinct sets of symptoms ranging from fever, joint and muscle pain to more severe manifestations like hemorrhage and neurological complications. Prognostic data outlines the typical progression and potential complications associated with each disease, guiding treatment strategies and patient management. Diagnostic parameters include laboratory tests, imaging findings, and geographical and temporal considerations crucial for disease confirmation and epidemiological analysis. This dataset serves as a foundational resource for developing machine learning models aimed at early disease detection, personalized treatment planning, and proactive public health interventions tailored to combat the dynamic and evolving challenges posed by vector-borne diseases globally.

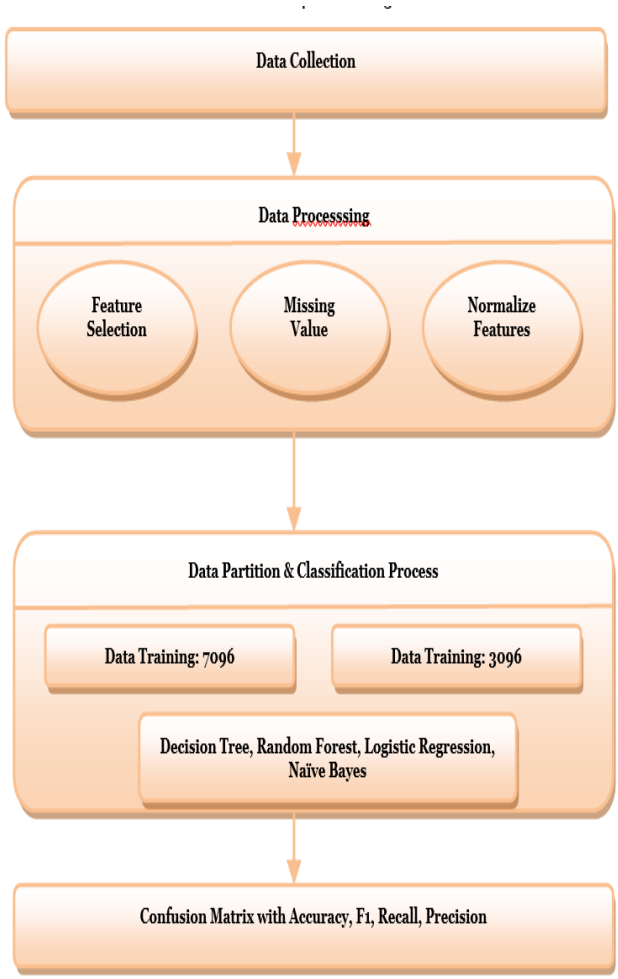

Figure 4. Methodology for development of prediction model for Dengue

The following step involves performing data preprocessing functions that both select features along with missing data elimination. Standardization takes place for the chosen features following scale normalization and threshold definition. Data partitioning and classification occurs as the last phase through systematic predictive model development. The data is separated into training and testing components by following a ratio of 70:30 for split purposes. The proposed prediction model uses machine learning classifiers like Decision Tree, Random Forest, Logistic Regression along with Naïve Bayes for its construction. In the last phase of the model development a confusion matrix enables performance assessment through precision together with F1 score recall and accuracy.

3. Machine learning prediction model

3.1. Prediction model for diagnosis

The proposed prediction model contains distinct modules that include Data Transformation followed by Data Preprocessing and then Dataset Partitioning and finally Model Development completed with Prediction. The Data Transformation module exists to retrieve stored data then restructure raw or semi-structured data into relational databases. The data warehouse provided CSV files or database query access which the author used for data extraction. The second module emphasizes record refinement while data normalization takes place based on the prediction model requirements. The preprocessing initiative stretches across all dengue fever characteristics by conducting proper scaling while also purifying data and removing outlier points. The resulting prediction model dataset achieves high quality through a well-defined systematic process which removes all errors. The third module of partitioning divides data between training and testing data by maintaining a 70:30 ratio. The prediction model develops based on training data until its evaluation phase uses test data. In the fourth module of the workflow teachers train the predictive model through an operation that utilizes both training and testing databases.

The team implements Gaussian Naïve Bayes Classifier within the prediction model after finishing these steps. External source information is processed by the Prediction Module through its designed system to create predictions. ThePrediction Module allows users to conduct early detection in addition to type-classification of dengue fever. This module serves two purposes including early detection of dengue through analyzing symptoms. The predictive system uses clinical reports for classification between minor and severe dengue cases. Patient records are collected by the prediction module through mobile applications and web applications as well as application software. There is a preprocessing step that follows patient data reception before the data moves to data cleaning. The system conducts patient record evaluations through either symptom-based analysis or symptom-based and clinical report analysis. The module forecasts dengue at its initial stages when the system receives only symptom data. Medical cases receive a confirmed dengue diagnosis status when symptom and clinical report data are combined in the input data. The module establishes both the normal or severe classification of dengue infection for every case. A previous explanation details the function of all modules incorporated in the vector-borne disease model.

3.2. Prediction model for vector-borne disease

The following section explains the complete operation of the prediction diagnostic framework while explaining its functions specifically for detecting vector-borne diseases starting with Dengue. The system incorporates five main functional blocks that start with data conversion then progress to data preparation before dividing the dataset between training and test sets for constructing a model and the final step is prediction. The system contains two fundamental data sets including historical data collection and newly obtained data. A local server functions as the repository for historical data storage together with its management system. The investigator selected this particular local server as their data platform. The storage system saves historical data in MySQL databases along with .csv file type. The information derives from predictions about dengue cases as well as records of rule derivations and raw data obtained from clinical laboratories. The information comes from mobile applications and web applications for users and from clinical laboratories. The next procedure requires extraction of features from existing records together with newly obtained data. The analysis utilizes descriptive statistics to make a connection assessment between different features. The imported dataset gets split into training and test sections on proportions of 70 and 30 percent respectively. The model passes an evaluation test during its development stage before undergoing result assessment. When development is complete a new data processing system with prediction capabilities becomes ready for deployment.

3.3. Data Transformation Module

This portion identifies the procedure for transforming raw data along with unstructured data into an analysis-ready format. Figure 5 shows how the eight-step process of data transformation takes place according to the procedure. Python pandas stores semi-structured data received from different storage sources in a new dataset.

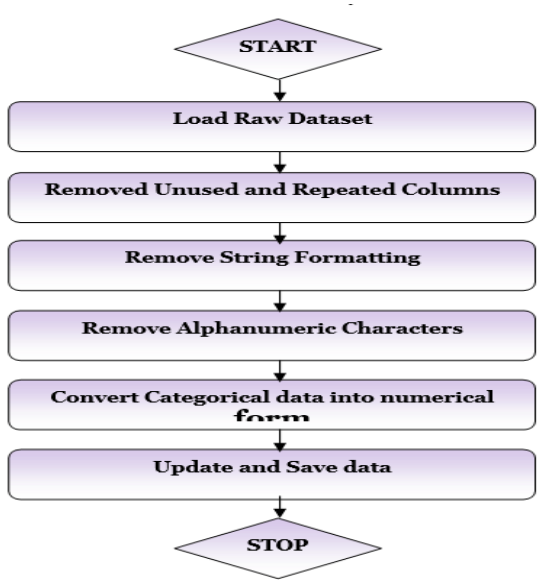

Figure 5. Data Transformation

Data cleanup demands specific methods to eliminate useless columns that exist in the dataset during its second procedural step. The third step of data cleaning processes removes string formatting along with alphanumeric characters from the dataset. String data gets converted into readable text which gets saved into proper columns of the dataset during this phase. The cleaned data is prepared for additional analysis after finishing its purification process.

During step 4 and for analysis purposes the transformation process moves forward by converting all categorical data into numerical values. The conversion takes place through scales that specify the particular function requirements. The prediction model contains specific encoding measurement scales to convert categorical data attributes including fever levels with values of high, medium, low and zero.

The timestamp data undergoes correct format conversion in the fifth step. The modified data completes storage in a dataset for preprocessing and normalization while Figure 5 illustrates this process (step 6).

Algorithm

Raw Data : D_{raw}
Regex Extraction: E(D_{raw})
//Function to apply regex to extract relevant data fields.
Normalization: N(E(D_{raw}))
// Function to normalize extracted data fields)
Structuring: S(N(E(D_{raw})))
// Function to structure normalized data into a semi-structured format)

Multiple essential actions must be taken to convert raw data records into a semi-structured vector-borne disease dataset. Regular expressions E(D_raw) work to parse and arrange raw data D_raw in order to extract diagnostic criteria and symptoms as well as geographic information. The extracted data passes through normalization N(E(D_raw)) which standardizes information while proper formatting dates and unit measurements must be unified when possible. The normalized data moves through the structuring process S(N(E(D_raw))) by becoming arranged into standardized formats which often use JSON as an example. A formatted approach allows for systematic data analysis and makes it possible to use machine learning algorithms for vector-borne disease forecasting and management techniques. The transformation process comprises essential phases which establish a foundation for improving data usage effectiveness and quality as a groundwork for healthcare decisions and preventive treatment.

3.4. Data preprocessing, function scaling and normalization module and Feature Selection

The predictive model depends on this section where two critical processes are conducted: feature engineering and data preprocessing. Primary features from the dataset are extracted in the feature engineering module's first phase which enables establishing inference rules for the planned predictive model as Figure 6 illustrates. At the start of the process the transformation module executes the loading of the dataset. Significant features related to early detection and classification of dengue are recognized during this stage. The researcher establishes diagnostic rules to identify dengue disease in the next stage. After an analysis of selected features' relationships the prediction model retains those features which demonstrate dependency. During this step the elements are resized before data values get converted to the necessary format and data type.

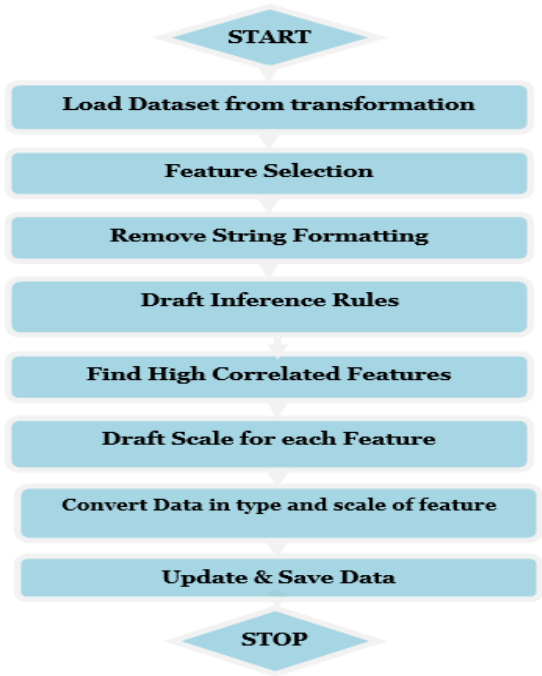

Figure 6. Data Preprocessing and Feature Scaling

3.5. Data pre-processing:

At the beginning of the process the transformation module operates to fetch the dataset. The database provides data retrieval after selecting vital features in the second stage of data processing and cleaning. The researcher proceeds to examine and address absent data values and replace them with correct elements in the third processing step. Identifying and eliminating duplicate records stands as the fourth step while the fifth step requires outlier detection to remove them. The sixth step of normalization operates on every dataset component to maintain uniformity. All modifications within this section get safely stored on the server at the conclusion

of the seventh step. Normalization and Min-Max Scaling and standardization are part of the preprocessing stage.

5.5.1 Min- Max Scaling

The Min-Max technique applied to functions transforms variable values into a standardized 0 to 1 range by using minimum and maximum values found in the input data. All feature values get proportionally modified through this process to match the range from 0 to 1. The following formula represents the Min-Max Scaling formula:

$$\frac{x-\min(x)}{\overline{x}=\max(x)-\min(x)}$$

3.5.2 Standardization

Standardization transforms data features through an adjustment process which sets their central value (or mean value) at zero and applies a standard deviation value of one. The normalization method known as Z-score produces standardized values which are written as "z" in the output. The standardized method will maintain an equivalent data distribution skew that exists in the original data set.

$$\frac{z = x - \mu}{\sigma}$$

Figure 7. Normalization

3.5.3 Normalization:

Normalization expands all vectors in the coordinate space while shrinking or lengthening their lengths. Every feature selected during feature selection and correlation matrix analysis has been chosen by the researcher in this study. The Dataset needs this process for proper creation. Scaling features accurately ensures that the database formation remains precise. Features that undergo scaling help predict models avoid errors while improving their final performance standards. The next process demands the researcher to analyze both the scale and normalized values of features while focusing specifically on binary classification in the research. The investigator adjusts the scales of patient features particularly fever variables to use zero as the minimum value and one as the maximum.

3.6. Divide dataset module:

The preprocessed dataset was divided into three distinct sets to ensure a comprehensive evaluation of the model's performance. Specifically, 70% of the data was allocated for training purposes, allowing the model to learn from a substantial portion of the dataset. An additional 15% of the data was designated for testing, providing an initial measure of the model's accuracy and effectiveness [19]. The validation segment occupied 15% of the data to conduct final assessments which refined the model following its application to unforeseen data. The organized methodology provides reliable data assessments throughout all stages of testing [20].

The Gaussian Naïve Bayes Classification machine learning algorithm powers the model construction module which enables the completion of the envisioned diagram shown in Figure 7.

3.7. Model building module

The Naive Bayes model uses probabilistic foundations together with an assumption of independent features to identify vector-borne diseases. Initial disease probability discovery occurs by analyzing historical datasets that contain symptoms together with region data alongside patient demographics. Training enables the model to determine both feature likelihood when assigned to each class along with initial probabilities of class occurrence. New data input leads the model to apply Bayes' theorem for calculating the posterior class probabilities based on the observed features. The model performs its classification by estimating the predicted disease from the most probable class. Despite its simplifying assumption of feature independence, Naive Bayes can effectively categorize vector-borne diseases by utilizing these probabilistic calculations to make informed classifications

based on the data's patterns and distributions [21].

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_{2,\ldots\ldots}, x_n)}$$

Where, $P(y \mid x_1, x_2, \ldots, x_n)$ is the posterior probability of class y given features $x_1, x_2, \ldots, x_n$, P(y) is the prior probability of class y, $P(x_i|y)$ is the likelihood of feature x_i given class y. $P(x_1, x_{2,\ldots\ldots}, x_n)$ is the evidence, which acts as a normalizing constant.

3.5.4 Feature Selection:

Feature selection is the process of identifying and selecting the most relevant features (variables) from a dataset to improve model performance by reducing overfitting and enhancing interpretability**.**

Information Gain (IG): Measures the reduction in entropy or uncertainty about the class label C when a feature x_i is known:

$$IG(x_i, C) = H(C) - H(C|x_i)$$

where H(C) is the entropy of the class labels and $H(C|x_i)$ is the conditional entropy of C given (x_i).

Chi-Squared Test: Assesses the independence between each feature (x_i and the class C. The chi-squared statistic is calculated as:

$$x^2(x_i, C) = \sum_{x_i \in values((x_i)} \frac{(O_{x_i} - E_{x_i})^2}{E_{x_i}}$$

where O_{x_i} is the observed frequency of class C for feature x_i, and E_{x_i} is the expected frequency under independence.

Recursive Feature Elimination: Recursive Feature Elimination (RFE) can be applied to Naive Bayes by recursively training the classifier and removing the least important features based on model coefficients or feature ranking scores

1.Initialization: Start with all p features in the dataset X=$\{X_1, X_2, \ldots X_p\}$.

2. Train Machine Learning Model on the current set of features X. Obtain feature important scores from model.

3. Rank the features based on their importance scores.

4. Eliminate the least important feature(s) from X based on the ranking. Update X to exclude the eliminated features.

5. Repeat steps 2-4 iteratively until the desired number of features is selected or a stopping criterion (e.g., a predefined number of features, no improvement in model performance) is met

Regularized Naive Bayes:

These techniques typically involve adding regularization terms to the Naive Bayes model to penalize certain feature coefficients, thereby influencing feature selection

Objective Function: The objective function typically includes a regularization term λ added to the log-likelihood or error function, aiming to penalize complex models or features that may overfit:

Objective = Log-Likelihood + $\boldsymbol{\lambda}$. Penality

Common penalty terms include L1L1L1 (Lasso) or L2L2L2 (Ridge) norms applied to feature coefficients or priors:

L1 Regularization (Lasso):

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

L2 Regularization (Ridge):

$$\lambda \sum_{j=1}^{p} |\beta_j^2$$

Where β_j are again the coefficients of features X_j

During model training, the regularization term λ is tuned to balance between fitting the data well and keeping the model simple.

Impact: Regularization encourages sparse solutions by shrinking less important feature coefficients towards zero, effectively performing feature selection.

The selection criteria for features relies on the coefficients which remain nonzero after regularization to identify important factors for model operation.

3.8. Prediction module

The chapter describes the methods for vector-borne dengue forecasting through the developed prediction model. The prediction model provides diagnosis for two different types of dengue patients who either do or do not report their symptoms. The model for prediction will serve healthcare providers by evaluating data that patients or laboratory technicians feed into its system. The first phase of data collection and examination allows patients who do not present NS1 or IgM reports to share their records for dengue fever diagnosis. Subsequent subsections detail the techniques and methods for diagnosing dengue fever in this context. Healthcare professionals including patients and medical staff present records with NS1 test results and blood antibody levels (IgM, IgG) together with platelet levels to patients.

The prediction system efficiently identifies confirmed dengue fever types and stages through a process covering both initial diagnostic and classification procedures.

Classification: Given new data instances x_{new}, the classifier predicts the most probable class $\hat{y}$ using Bayes' theorem:

$$\hat{y} = arg\ max\ P(y) \prod_{i=1}^{n} P(x_i|y)$$

The model determines class probabilities then predicts the class having the maximum probability as y^. Theresearcher presents a complete description of their work to develop a machine learning-driven vector-borne dengue disease prediction system which focuses on Indian stage classification and early diagnosis methods. The model comprises five modules: Data Transformation, Data Preprocessing, Feature Scaling and Normalization, Dataset Partitioning, Model Building, and Prediction. The first module executes multiple data transformation steps which start with raw data loading followed by redundant column removal and string formatting elimination and alphanumeric character clearing alongside categorical data conversion to numbers and timestamp adjustment.

The second module encompasses data preprocessing, feature scaling, and dataset normalization. Feature scaling comprises the content of module three while the fourth module describes preparing a proposed Gaussian Naïve Bayes classification model. Dengue-infected patient diagnostic tool uses a modified classification approach to speed up dengue detection and diagnostic stage identification. The researcher built four algorithms which identify dengue exposure during its early stages with the categories including Dengue-DF for normal dengue and the subsequent classifications of Dengue-DHF and Dengue-SD for severe stages. The current stage of development enables the proposed vector-borne disease prediction model to diagnose dengue infections at different stages prior to taking it through tests and validation steps using real-time data.

The subsequent chapter delves into the testing and validation of this proposed model for vector-borne dengue disease.

The inference rule contains seven vital components needed to identify dengue early (Figure 8) that follow the 2009 WHO guidelines. The 2009 WHO guidelines state that two simultaneous symptoms among fever, nausea, rash, abdominal pain, vomiting and retroorbital pain should trigger dengue infection assessment. Each dengue symptom points toward an average of six other symptoms according to the correlation matrix evaluation. This data displays co-relationships between rash related to nausea together with vomiting being connected to pain thus demonstrating a strong probability for such connections. Patients with fever along with pain tend to develop nausea and rash and vomiting according to the correlation matrix findings.

In machine learning, a property refers to a variable or characteristic of an object. Every element receives a numerical value that spans from maximum to minimum possible limits. The Element Scale describes data intake for specific elements through the established maximum and minimum numerical values. The model responds according to the input values assigned to its elements within this framework. The procedure of defining the boundaries for an element falls under normalization. Between standardization and normalization and minimum-maximum scaling machine learning uses these practices most frequently as feature scaling approaches. The table provides additional information about these methods as shown in Table 2.

A person's excessive body temperature serves as the key

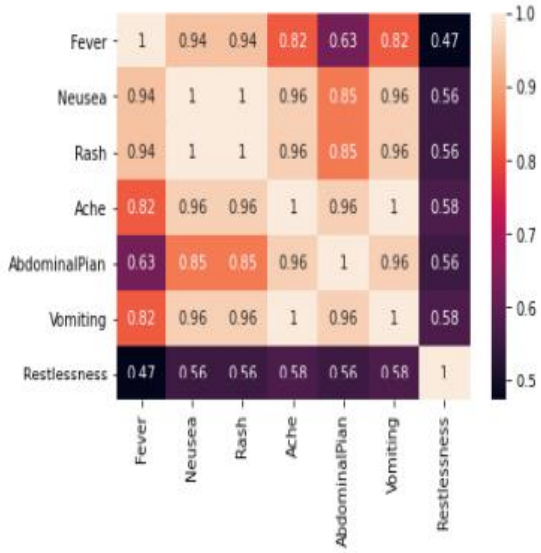

Figure 8. Correlation Matrix of Inference Rule

indicator of dengue disease while other symptoms like headache and nausea and vomiting and retroorbital pain and joint pain appear during the infection period. The researcher established a classification system for fever which divides the temperature condition into four categories: Normal, Low, Medium, and High for prediction development purposes. The index value gets established from data points connected to fever input variables described in Table 2.

Performance Metrics

Performance metrics are essential in evaluating and refining machine learning models. They provide quantitative measures to assess model effectiveness, highlight weaknesses, and guide improvements

Accuracy: The ration of correctly predicted instances to the total instances.

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision (P): Ratio of correctly predicted positive observations to the total predicted positives.

$$P = \frac{TP}{TP+FP}$$

Recall (R): Ratio of correctly predicted positive observations to all the observations in the actual class.

$$R = \frac{TP}{TP+FN}$$

F1-Score:The weighted Averages of Precision and Recall.

$$F_s = 2 * \frac{Precision*Recall}{Precision+Recall}$$

Table 2. Measuring Table

Measuring table					
Function	**Type**	**Parameters**	**Index**	**Normalized**	
Rectal	Normal	97.9°F-100°For 36.6°C-38 °C	0	0	
Oral	Normal	95.5°F-99.5°For 35.5°C-37.4 °C	0		
Axillary	Normal	94.5°F-99.3°For 34.7°C-37.4 °C	0		
Tympanic	Normal	96.3°F-100°For 34.7°CC	0		
Temporal	Normal	97.9°F-100.1°For36.6°C-37.8°C	0		
Oral	Fever	>99.5° for>37.5°C	1	1	
Rectal	Fever	>100.4° for37.5°C	1		
Axillary	Fever	>100° for37.8 °C	1		
Infant	Fever for Patient 0to 2	>100.7 for38.1°C	1		
	months				
Baby	Fever for patient 3 to 47 months	>100.3 for37.9°C	1		
Child	Fever for patient older than 4 years	>100.1 for37.8°C	1		

These metrics facilitate clear communication of model performance to stakeholders, ensuring that the chosen model aligns with the project's goals and constraints.

5. Result and Discussion

The study investigates the validation approaches that measure the effectiveness of the proposed vector-borne disease prediction method. The research paper demonstrates through a systematic approach each step required for validating and testing the developed model. The research methodology proceeds through two distinct phases where it evaluates the model using different machine learning classifiers before measuring dengue case accuracy in the proposed model framework.

Table 3 performance of proposed model with Biosensor

	Accuracy	Precision	Recall	F1-Score
Navie bayes	97.39%	96.58%	97%	95.98%

The performance metrics derived from the Naive Bayes classifier for predicting vector-borne diseases shown in Table 3 and Figure 9, illustrate its effectiveness in classification tasks. Based on the observed features of the dataset the model shows an accuracy rate of 97.39% in its disease classification process. The model achieves precision at 96.58% through the measurement of true positive predictions among its total positive predictions ensuring very low numbers of false disease identification. The model shows capability to detect most actual positive cases in the dataset with a recall rate of 97%. A single metric F1-score reveals a percentage of 95.98% which demonstrates optimal performance between precision and recall in disease prediction systems. The Naive Bayes classifier stands as a reliable solution for obtaining effective and precise vector-borne diseases predictions because of its performance metrics.

Figure 9 Performance of proposed model

5.1. Proposed model Performance measure for vector borne diseases

The section focuses on presenting results from testing and validating the prediction model through multiple machine learning techniques and algorithms. The model testing includes data comparisons with five different machine learning algorithms according to the information in Table 3.

Naive Bayes obtains superior performance to alternative methods by applying its efficient probabilistic approach which is shown in Table 4 and Figure 10. Naive Bayes works efficiently using an independence assumption that handles datasets with numerous features. Naive Bayes fits big data sets with ease through its basic computational structure which needs few training resources and processing time. Naive Bayes demonstrates good results in conditions that satisfy the independence assumption either exactly or when

features become conditionally independent of the class label. Naive Bayes is also robust to noise and irrelevant features, as it can effectively ignore non-informative attributes during classification. These factors contribute to its ability to provide accurate predictions swiftly, making it ideal for real-time applications such as spam detection or text classification. However, its performance can vary based on the dataset's characteristics and the strength of feature dependencies.

Table 4 Performance Comparison of Proposed Model for Vector-Borne Diseases

ML	%	Cross Validation	K-fold Validation	Pr	Recall	F1	R-Squared	MSE
Decision Tree	97.5%	0.968	0.968	0.98	0.97	0.97	0.46	0.03
Logistic Regression	97.5%	0.975	0.975	0.98	0.97	0.97	0.46	0.03
SVM	96.87%	0.975	0.97	0.98	0.97	0.97	0.46	0.03
Proposed Model	97.39%	0.97	0.973	0.98	0.975	0.97	0.55	0.02

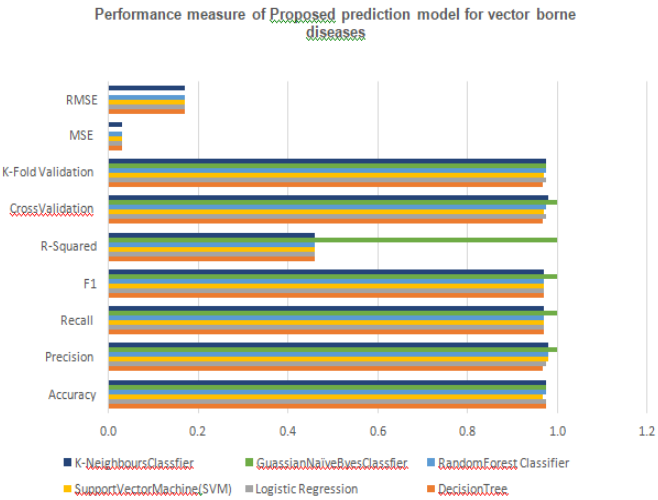

Figure. 10. Proposed model Performance measure for vector borne diseases

CONCLUSION

The identification of vector-borne diseases through biosensors creates substantial annual challenges for countries both in advanced and underdeveloped regions with special significance in India, Bangladesh, and Sri Lanka. These countries spend a large share of their budget on disease management while also actively working to prevent such diseases which negatively impacts public health. Indian authorities have struggled with the same health issue since 1970 as their society experiences different patterns in geography together with economic variations and cultural and social shifts. Healthcare professionals must take an active approach to medical delivery through Internet and information technology integration because the current passive healthcare system requires improvement. The conducted research helps develop a system which predicts dengue early detection alongside its classification and identification of serious conditions. The research follows sequential stages starting with literature review and problem identification before defining methodology and designing the prediction model and implementing it for analysis of experimental results. These four essential elements make up the dengue prediction model according to the researcher: dengue outbreak forecasting, disease detection, correct classification of types and the identification of severe patient cases. Research on dengue detection focuses either on identifying outbreaks at locations or detecting single cases with additional interest in classifying disease severity. The researcher takes on the task of creating a prediction model to detect early dengue cases and classify their symptoms based on clinical presentations since this area remains underserved. Through this model entities can identify dengue cases early and classify different dengue types by processing clinical patient information. Rigorous feature selection allowed the researcher to establish frequency relationships between dengue-related symptoms before creating inference rules and determining diagnostic correlations for determining essential diagnostic indicators. The designated algorithm achieves successful testing on the Python platform following its design and implementation. Five components within the proposed model show their ability to diagnose dengue disease at different points in its progression. Testing and validation of the model led to identification of 97.3% accuracy in the subsequent phase. Different machine learning algorithms demonstrate how versatile the model is and the decision tree and Gaussian Naive Bayes Classifier algorithms showcase high accuracy and reliability.

Evaluating the findings and real-world impacts requires an analysis of both practical benefits and application effects of using Naive Bayes to forecast vector-borne diseases. Naive Bayes classifier proves valuable for detecting diseases early because it simplifies the analysis of extensive datasets to help monitor malaria, dengue, and Zika illness. History-based disease outbreaks provide an excellent basis for the model to identify future outbreak patterns through accurate prediction of environmental and socio-economic factors. The prediction model allows health authorities to direct their resources effectively by enabling them to plan focused prevention methods that stop disease spread. Administrative health

surveillance programs gain improved effectiveness when they implement predictive models which enhances their ability to support medical readiness and lessen disease impact on communities. The research outcomes demonstrate how machine learning especially Naive Bayes algorithms offer significant transformative benefits for disease prognoses and public healthcare operations during vector-borne disease protection routines.

REFERENCES

[1] K. P. Murphy, Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press, 2012.

[2] H. Wang, C. Ma, and L. Zhou. "A Brief Review of Machine Learning and Its Application," in 2009 International Conference on Information Engineering and Computer Science, Dec. 2009, pp. 1–4. doi: 10.1109/ICIECS.2009.5362936.

[3] NVBDCP. (2021, Jun, 11). Dengue :: National Vector Borne Disease Control Programme (NVBDCP) [Online] Available: https://nvbdcp.gov.in/index1.php?lang=1&level=1&sublinkid=5776&lid=3690

[4] Rajesh, T.R., Rajendran, S. and Alharbi, M., 2023. Penguin search optimization algorithm with multi-agent reinforcement learning for disease prediction and recommendation model. Journal of Intelligent & Fuzzy Systems, 44(5), pp.8521-8533. DOI: 10.3233/JIFS-223933

[5] Logapriya, E. and Surendran, R., 2023, November. Hybrid Recommendations System for Women Health Nutrition at Menstruation Cycle. In 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 357-363). IEEE. doi: 10.1109/3ICT60104.2023.10391518.

[6] Somasundaram, R., Selvaraj, A., Rajakumar, A. and Rajendran, S., 2023. Hyper Spectral Imaging and Optimized Neural Networks for Early Detection of Grapevine Viral Disease. Traitement du Signal, 40(5). DOI: https://doi.org/10.18280/ts.400528

[7] Gunasekar, G., Krishnamurthy, A., Thanarajan, T. and Rajendran, S., 2023. SFoG-RPI: A Secured QoS Aware and Load Balancing Framework for FoG Computing in Healthcare Paradigm. Revue d'Intelligence Artificielle, 37(4). DOI: https://doi.org/10.18280/ria.370403

[8] Jeyabharathi, M., Jayanthi, K., Kumutha, D. and Surendran, R., 2022, September. A Compact Meander Infused (CMI) MIMO Antenna for 5G Wireless Communication. In 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 281-287). IEEE. DOI: 10.1109/ICIRCA54612.2022.9985730

[9 Thanarajan, T., Alotaibi, Y., Rajendran, S. and Nagappan, K., 2023. Eye-Tracking Based Autism Spectrum Disorder Diagnosis Using Chaotic Butterfly Optimization with Deep Learning Model. Computers, Materials & Continua, 76(2). https://doi.org/10.32604/cmc.2023.039644

[10] Dourjoy, S.M.K., Rafi, A.M.G.R., Tumpa, Z.N. and 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). Springer Singapore. DOI: 10.1007/978-981-15-4218-3_49

[11] Sangeetha, T., Kumutha, D., Bharathi, M.D. and Surendran, R., 2022. Smart mattress integrated with pressure sensor and IoT functions for sleep apnea detection. Measurement: Sensors, 24, p.100450. https://doi.org/10.1016/j.measen.2022.100450

[12] A. Pravin, T. P. Jacob, and G. Nagarajan, "An intelligent and secure healthcare framework for the prediction and prevention of Dengue virus outbreak using fog computing," Health and Technology, vol. 10, no. 1, pp. 303–311, Jan. 2020, doi: 10.1007/s12553-019-00308-5. 161

[13] S. K. Sood, S. Kaur, and K. K. Chahal, "An intelligent framework for monitoring dengue fever risk using LDA-ANFIS," Journal of Ambient Intelligence and Smart Environments, vol. 12, no. 1, pp. 5–20, 2020, doi: 10.3233/AIS-200547.

[14] N. Siringi, S. Mala, and A. Rawat, "Study of K-Means Clustering Algorithm for Identification of Dengue Fever Hotspots," in ICDSMLA 2019, Singapore, 2020, pp. 51–61. doi: 10.1007/978- 981-15-1420-3_6.

[15] Shubham Kumar 1, Aman Srivastava 1 2, Rajib Maity, "Modeling climate change impacts on vector-borne disease using machine learning models: Case study of Visceral leishmaniasis (Kala-azar) from Indian state of Bihar", Expert Systems with Applications 2024, vol:237. https://doi.org/10.1016/j.eswa.2023.121490

[16] Daozhou Gao & Linlin Cao, "Vector-borne disease models with Lagrangian approach", Journal of Mathematical Biology, 2024, vol:88. https://doi.org/10.1007/s00285-023-02044-x

[17] Ritesh Chandra, Sadhana Tiwari, Sonali Agarwal, Navjot Singh, "Semantic web-based diagnosis and treatment of vector-borne diseases using SWRL rules", Knowledge-Based Systems, 2023, vol:274. https://doi.org/10.1016/j.knosys.2023.110645

[18] Inderpreet Kaur, Amanpreet Kaur Sandhu & Yogesh Kumar, "Artificial Intelligence Techniques for Predictive Modeling of Vector-Borne Diseases and its Pathogens: A Systematic Review", Archives of Computational Methods in Engineering, 2022, vol:29, pp:3741. https://doi.org/10.1007/s11831-022-09724-9

[19] Sundarapandi, A.M.S. and Rajendran, S., 2024. Sensors Based Optimized Closed Loop Control Algorithm to Minimize Hypoglycemia/Hyperglycemia using 4-Variate Time Series Data. Journal of New Materials for Electrochemical Systems, 27(1). https://doi.org/10.14447/jnmes.v27i1.a01

[20] Lv, Y., Yang, L., Bu, F. and Yang, J., 2023. Optimization of Sensor Array for Detection of Abalone Freshness Based on Electronic Tongue. Journal of New Materials for Electrochemical Systems, 26(1). https://doi.org/10.14447/jnmes.V26i1.a11

[21] https://www.kaggle.com/datasets/richardbernat/vector-borne-disease-prediction/data