# VECTOR BORNE DISEASE PREDICTOR

**by**

## M SHARATH SRIVATSAN 19BCE1688

**NAVIN THOMSY 19BCE1695**

**CHANDRU S 19BCE1878**

**AYAN SADHUKAN 19BCE1872**

A project report submitted to

# Dr. Maheswari R

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

in partial fulfilment of the requirements for the course of

# ECE3502 IoT Domain Analyst

in

# B. Tech. COMPUTER SCIENCE AND ENGINEERING



**Vandalur – Kelambakkam Road Chennai – 600127**

## APRIL 2022

## ABSTRACT

## Our country, with such a huge population and with the favorable geo-ecological setting is very highly susceptible to the impact caused by the Vector Borne Diseases (VBD). Considering the hygiene, lifestyles of people in our country and the rate at which diseases spread, there is a need to act upon the spread of VBDs immediately. Vector-borne diseases (VBDs) are sickness resulting from parasites, viruses and microorganism which can be transmitted through vectors. These VBDs include Malaria, Dengue, Chikungunya, Japanese Encephalitis (JE), Lymphatic Filariasis and Kyasanur Forest Disease (KFD). In this proposal, the term VBD refers to the most affecting diseases such as malaria and Dengue.

## Health information demands are also influencing information seeking behavior, as seen all across the world. Many people experience difficulties while seeking for health information online on ailments, diagnostics, and treatments. It will save a lot of time if a recommendation system for physicians and medicine can be developed utilizing review mining. Since the users are laymen, they have difficulty grasping the diverse medical jargon in this sort of system. As there is so much medical information available on many channels, the user gets perplexed. The concept behind the recommender system is to respond to the unique needs of the health sector in terms of users.

## INTRODUCTION

## It is a known fact that climate can have an effect on the way these VBDs spread i.e., on their transmission dynamics, geographic spread and their re-emergence. Tropical climates have proven to be favorable for the spread of these VBDs. Economically poor countries whose health services are not up to the mark are more vulnerable to these diseases i.e., developing nations are more susceptible to these diseases when compared with the developed nations. India with such a huge population and with the favorable geo-ecological setting is very highly susceptible to the impact caused by the Vector Borne Diseases. Considering the hygiene, lifestyles of people in our country and the rate at which diseases spread, there is a need to act upon the spread of VBDs immediately. Thus, the major challenge in the present condition is to trace and identify the patient location to provide the treatment. The different ways to prevent this communicable disease is by identifying the infectious persons and by providing effective and continuous treatment to those persons at the earliest in Government Primary/Public Health Centre in the rural region of our country.

## As Machine Learning, deep learning and IoT techniques are advancing very well, therefore these techniques can be used for analyzing, detecting, predicting, forecasting and even suggest changes which would help in prevention of the spread of these diseases.

## In the second part of the twentieth century, the worldwide human illness burden ascribed to vector-borne diseases (VBDs) grew dramatically. VBDs were projected to account for 17% of human illness burden in 2017, up 2.4 percent from 1990, albeit this figure is likely to be underestimated due to widespread under-reporting in many endemic countries. This surge has included the introduction and establishment of novel vector-borne infections (e.g., dengue, Zika, West Nile virus (WNV), chikungunya) as well as increased incidence and effects in endemic regions (e.g., Lyme disease, tick-borne encephalitis). The recent spread and epidemics of these illnesses have been linked to a number of interconnected social, ecological, and environmental variables.

## Rising temperatures, for example, have an impact on vector biting, survival, and reproduction rates, as well as pathogen growth and survival. Furthermore, shifting precipitation patterns have a variety of effects on vector breeding areas. To influence the transmission and development of VBDs, these climatic factors interact with non-climatic variables such as globalization and urbanization, sociodemographic, and public health systems.

## LITERATURE REVIEW

Muthoni Masinde’s [1] research show that there is a very significant relationship between certain environmental and climatic factors and the incidence of these VBDs. Climatic conditions influence the life cycle and development of these vectors for instance, their maturity, biting capacity and survival periods. The author state that out of all other climatic variables temperature, humidity and rainfall affects the incidence of VBDs the most. Pollution and global vegetation index also significantly affect the lifecycle of the vectors. Temperature within the 20-30℃ range is said to be the optimum temperature for the survival of these vectors. Anything below 16℃ or above 35℃ inhibits their survival rates. There seems to be no straightforward relationship between the annual rainfall index and the vector, but in association with the other factors, there is some noticeable correlation. High humidity favors the survival of these vectors.

Gadson Kalipe [3] et al, in their study of predicting VBD outbreak, have used and presented a comparative analysis on the performances of different algorithms and approaches used for predicting. The authors have concluded that XGBoost, ANN, Random Forest and SVM outperforms other prediction models.

From prevention perspective, Wilson Alobuia and Pauline [4] presented a study, which calculates the KAP scores of individuals for these VBDs and show that regions, where the average KAP score is lesser than a said threshold, are more vulnerable to outbreaks than those regions with higher average KAP scores.

The different mathematical models like basic models, a hierarchical structure of a range of deterministic models of various levels of complexity for epidemiological diseases were discussed in their work [6].

Encryption keys and digital signature keys stored on smart cards of the system are used for secure and authenticated data communication between clients and database servers over distributed object protocol. The Taiwan health care smart card project is one of the largest health care smart card solutions in the world and the first of its kind in Taiwan, Republic of China. Before the smart card was introduced, paper cards were used by the Bureau of National Health Insurance (BNHI) to audit patient information, then reimburse service providers monthly [7]. The card is renewed after the patient uses medical services up to six times. Even though reporting and information handling is well run and maintained, the system has certain problems, such as identity fraud, excess false insurance premium claims from health care institutions, complex program vouchers, waste of resources due to high frequency of card replacement, and high losses due to discontinuity of insured applicants. To solve these problems, in April 2001 the BNHI issued 22 million smart health care cards using Java Card technology to Taiwanese citizens Dongguan et.al.,[8].

## EXISTING WORK

## Disease surveillance is used widely in both statistical and process-based infectious disease forecasting methodologies. Prior to making a forecast, observations of previous and current disease activity are utilized to improve model systems, which is crucial for delivering accurate and calibrated forecasts (Shaman et,2013). From this perspective, an ideally vector-borne illness monitoring system would contain frequent, high-quality observations of disease dynamics in both human and vector populations. In the human population, measurements would include prevalence and incidence, as well as immunological status, and in the vector population, rates of infection, lifespan, . Current vector-borne illness surveillance methods, on the other hand, are typically confined to passive monitoring of human infections (Beatty et al., 2010). These incomplete observations reduce the

## accuracy of vector-borne disease forecasting (Yámana et al., 2016)

## Few, however, have attempted to evaluate the utility of surveillance data in forecasting illness risk in the future. Kilpatrick and Pape (2013) investigated the value of mosquito infection rates as predictors in West Nile Virus regression models at various temporal and geographical scales. Davis et al. (2017) forecasted West Nile Virus using environmental variables and mosquito infection rates, noting an improvement in forecasting after mosquito data became available, but not comparing predictions with and without the data.

## Adult vector infection rates have recently been incorporated into disease surveillance systems to improve understanding of the spatiotemporal risk of disease transmission have used mosquito monitoring data to develop early warning or prediction systems for disease outbreaks.

## PROPOSED SYSTEM

## We want to build a Vector Borne Disease Prediction system that will accept any user's name and any probable symptoms of a vector borne disease as input. Since the number of vector-borne illnesses is limited, identifying symptoms for all of them was not difficult. We employ pre-existing sophisticated Machine learning models to identify which vector-borne disease the user is likely afflicted with after taking at least two and up to five symptoms from the user. As previously stated, we will be building a user-friendly GUI for all laymen, even those who are technologically challenged, in order to broaden the reach of our prediction tool.

## Every year, as the number of patients and diseases increases, the medical system becomes overburdened and, in many nations, expensive. The majority of the condition necessitates a visit with a doctor in order to be treated. With enough data, illness prediction using an algorithm can be simple and inexpensive. Predicting illness based on symptoms is an important element of treatment. In our study, we attempted to properly forecast an illness based on the patient's symptoms. For this, we utilized four different algorithms and achieved a 92-95 percent accuracy. A system like this has a lot of potential in future medical therapy. We've also created an interactive interface to make working with the system easier. We've also tried to illustrate and depict the findings of our research and this project.

## FLOW CHART OF PROPOSED WORK

## 

## WORKING MODULES AND DESCRIPTION

## Importing Libraries and Dataset

## Preprocessing and Visualization of Data

## ML Model(s) Implementation and Prediction

## Setting up Database

## Building the Graphical User Interface

## Importing Libraries and Dataset

## Dataset for this project was collected from Kaggle which was sourced by a study of university of Columbia performed at New York Presbyterian Hospital during 2004. Link of dataset is given below:

## <https://www.kaggle.com/datasets/richardbernat/vector-borne-disease-prediction>

## 

## Standard libraries for database analysis and model generation are implemented in this project. The libraries used in this project are listed below.

## tkinter: It’s a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It comes with a strong object-oriented tool for designing graphical user interfaces. It offers different widgets to implement a GUI. We were able to construct an interactive GUI for our model using tkinter.

## NumPy: NumPy is a Python package mainly used for scientific computing. In Python, it provides sophisticated capabilities for dealing with a variety of multi-dimensional arrays. It's a multi-purpose array processing programmed. The primary goal of NumPy is to work with multidimensional homogenous arrays. It offers tools for everything from array construction to array management. It makes it easy to build an n-dimensional array with simply np. zeros () and then manipulate its contents using methods like replace, arrange, random, save, and load. It also aids array processing using techniques such as total, mean, standard deviation, max, min, all, and so on. When employing operators like +, -, \*, and / or arrays built using NumPy, they behave differently from arrays created regularly. NumPy array has all of the properties and services listed above, making it ideal for our data-handling needs. Data manipulation in arrays while conducting various operations must produce the intended results, and output prediction necessitates such high operational skills.

## Pandas: It is the most widely used Python data analysis package. With back-end source code written entirely in C or Python, it gives highly efficient performance.

## Python data may be analyzed in two ways:

## Data frame

## Series

## In pandas, a series is a one-dimensional array that may be used to hold any data type.

## In Python, data frames are a two-dimensional data structure that stores data in rows and columns.

## In this project, the Pandas data frame is utilized extensively to leverage datasets for training and testing algorithms. Data frames make working with characteristics and outcomes much easier. Several of its built-in functions, such as replace, were employed for data modification and preprocessing in our research.

## Sklearn: Sklearn is a Python package that implements machine learning, pre-processing, cross-validation, and visualization methods. It includes a number of basic and effective data mining and data processing tools. Support vector machine, random forest classifier, decision tree, gaussian nave-Bayes, and KNN are only a few of the classification, regression, and clustering algorithms included. We utilized sklearn in this project to take use of built-in classification techniques such as decision trees, random forest classifiers, KNN, and naive Bayes. We also employed cross validation and visualization elements such as the classification report, confusion matrix, and accuracy score that were integrated into the software.

## 2.Preprocessing and Visualization of Data

## We then declare two arrays each for diseases and all possible vector borne symptoms of these diseases.

## 

## Since we are in the data pre processing stage, we will visualize the head of our training dataset to have some insights about the information in the dataset.

## Now, we visualize the prognosis pie data and also get a look into the correlation matrix between our symptoms and the diseases.

## 

## We also show the per column distribution graph and scatter matrix of our dataset for better insight and to get to know how each symptom weighs in proportion to the importance. Here below are the code snippets for the same

## 

## *Per Column Distribution Graph*

## Scatter Matrix

## Obtaining the training data from Kaggle, we process and visualize it in our Jupyter Notebook

## 

## 3. ML Model(s) Implementation and Prediction

## Now as stated, we have four different Machine learning models present in our project to predict the vector borne diseases which are:

## • KNN

## • Decision Tree

## • Random Forest Tree

## • Naïve Bayes

## A Decision tree is a categorization approach that is both successful and adaptable. It's utilized for visual recognition and categorization. Because of its versatility, it is utilized for categorization in exceedingly complicated issues. It's also capable of dealing with higher-dimensional challenges. It is made up of three parts: root, nodes, and leaves. The root of the tree contains the attribute that has the most impact on the result, the leaf verifies the value of a certain attribute, and the leaf delivers the tree's output.

## 

## The first prediction approach we employed in our study was the decision tree. It provides us with a 94% accuracy rate.

## The Random Forest Method is a supervised learning algorithm that may be used to classify and predict data. This algorithm is made up of four essential steps:

## 1. It obtains random data selections from our dataset.

## 2. It builds decision trees for every selection and sample dataset chosen.

## 3. At the step we are in now every result will be consolidated and voted on.

## 4. At the end, the most voted prediction gets selected and will be presented.

## 

## We utilised a random forest classifier with 100 random samples in this research, and the outcome was 93.22% accurate.

## The K Nearest Neighbor algorithm is a supervised learning method. It is widely used in pattern recognition and data mining.

## It works by identifying a pattern in data that connects data to outcomes, and it becomes better at recognizing patterns with each repetition.

## 

## We were able to categorize our dataset with 92 percent accuracy using K Nearest Neighbor.

## The Naive Bayes algorithm is a collection of algorithms based on the nave Bayes theorem. They all follow the same principle: each pair of predictions is independent of the others. It also assumes that each characteristic contributes equally and independently to the prediction.

## 

## We employed the Naive Bayes algorithm in our project to get a 91% accurate forecast.

## 4. Setting Up Database

## 

## If a database named "database.db" does not exist, one will be created using the "sqlite3" database. Now "Decision Tree" table is made in the database.db for storing data for the decision tree algorithm, if it does not already exist, using the sqlite "CREATE" function. The "INSERT" func in sqlite is used for inserting values into the DecisionTree table.

## 5. Building Graphical User Interface

## 

## Code of GUI for setting the initial values

## The tkinter package in Python is used to create the graphical user interface. To launch the GUI, you must be logged in as root. The backdrop is set to "Ivory" in the configuration. title () function in tkinter library is named as "Vector Borne Disease Predictor System."

## Code for the GUI Message Box

## 

## Code for the GUI Option Menu

## 

## Code for the GUI Buttons

## 

## Code for the GUI Result Display

## SCREENSHOTS OF OUTPUT

## 

## *The Final Result Display*

## 

## *A label is used to create a heading and a section for contributors.*

## 

## *Labels are often used to distinguish distinct parts.*

## 

## *OptionMenu is a drop-down menu creator.*

## 

## *Buttons are used to provide functionality and anticipate the outcome of models, as well as two utility buttons, exit and rest.* *With the usage of blank space, text is employed to display the prediction's outcome.*

## 

## *Messageboxes are used in three distinct ways. One is for making name input mandatory*

## 

## Two – The need for at least two symptoms to be inputted.

## 

## three- to exit the GUI system.\

## 

## DB Browser for SQLite which stores User Info and Database

## 

## 

## Confusion Matrix and Accuracy of each model printed after prediction

## 

## *Complete Output*

## CONCLUSION

## We set out to develop a system that can forecast vector-borne illnesses based on the symptoms provided.

## A system like this might minimize the amount of time people spend in hospital emergency rooms and lessen the amount of work medical personnel has to do. We were successful in developing such a system, and we did it using four separate algorithms. We had a 96.2 percent accuracy rate on average. Such a system can be generally dependable in its performance. We also provided a means for the user's data to be saved in the database, which may be utilised in the future to aid in the development of a better version of the system. In addition, our system offers a user-friendly interface. It also includes a variety of visual representations of the data gathered and the outcomes obtained.

## FUTURE WORK

## We want to create a prediction model that takes into account environmental parameters in order to anticipate VBD outbreaks at a specific place based on its climatic conditions rather than just a patient's symptoms. The development of a Susceptible-Exposed-Infectious-Removed (SEIR) model to examine the spread speed of these illnesses is also possible. Furthermore, using QGIS, it is possible to see VBD locations. Importantly, a Knowledge, Attitudes, and Practices (KAP) research might be conducted to measure awareness, comprehension, and to identify locations based on vulnerability by calculating an average KAP score.

## REFERENCES

[1] *Muthoni Masinde*, Africa's Malaria Epidemic Predictor: Application of Machine Learning on Malaria Incidence and Climate Data, *ICCDA 2020: 2020 The 4th International Conference on Compute and Data Analysis*

[2] *Wu, Yan & Lee, Gary & Fu, Xiuju & Hung, Terence Gih Guang. (2008).* Detect Climatic Factors Contributing to Dengue Outbreak based on Wavelet, Support Vector Machines and Genetic Algorithm. *Lecture Notes in Engineering and Computer Science.*

[3] Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach. *Kalipe, Godson & Gautham, Vikas & Behera, Rajat. (2018). A Review and Analysis.*

*33-38. 10.1109/ICIT.2018.00019.*

[4] Knowledge, Attitude, and Practices Regarding Vector-borne Diseases in Western Jamaica. *Alobuia, Wilson & Missikpode, Celestin & Aung, Maung & Jolly, Pauline. (2015). Annals of Global Health. 81. 654-663. 10.1016/j.aogh.2015.08.013.*

[5] Shah, N. H., & Gupta, J. (2013)." SEIR model and simulation for vector borne diseases".

[6] Mandal, S., Sarkar, R. R., & Sinha, S. (2011). Mathematical models of malaria-a review. Malaria journal, 10(1), 1-19.

[7] The Taiwan Health Care Smart Card Project- smart card alliance Yuh-Ning Chen, Ph.D., MartSoft Corporation.

[8] Design and implementation of Smart Card COS Chen Yuqiang; Dept. of Computer. Eng., DongGuan Polytech., Dongguan, China; Guo Jianlan; Hu Xuanzi; Liu Liang, Computer Application and System Modeling (ICCASM), 2010 International Conference on (Volume:13)

[9] S. Raizada, S. Mala and A. Shankar, "Vector Borne Disease Outbreak Prediction by Machine Learning," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 213-218, doi: 10.1109/ICSTCEE49637.2020.9277286.

[10] https://royalsocietypublishing.org/author/White%2C+Steven+M

[11] Arnold and Dey, 1986 C.P. Arnold Jr, C.H. Dey **Observing-systems simulation experiments: past, present, and future** Bull. Am. Meteorology. Soc., 67 (6) (1986), pp. 687-695

[12] H. Auerswald, L. Klepsch, S. Schreiber, J. Hülsemann, K. Franzke, S. Kann, *et al.*

The dengue ED3 dot assay, a novel serological test for the detection of dengue virus type-specific antibodies and its application in a retrospective seroprevalence study

Viruses., 11 (4) (2019), p. 304

[13] P. Barbazan, S. Yoksan, J.-P. Gonzalez Dengue haemorrhagic fever epidemiology in Thailand: description and forecasting of epidemics Microbes Infect., 4 (7) (2002), pp. 699-705.

[14] M.E. Beatty, A. Stone, D.W. Fitzsimons, J.N. Hanna, S.K. Lam, S. Vong, *et al.*

Best practices in dengue surveillance: a report from the Asia-Pacific and Americas Dengue Prevention Boards PLoS Negl. Trop. Dis., 4 (11) (2010), p. e890

[15] L.R. Bowman, S. Runge-Ranzinger, P. McCall Assessing the relationship between vector indices and dengue transmission: a systematic review of the evidence

PLoS Negl. Trop. Dis., 8 (5) (2014), p. e2848

**COMPLETE PROGRAM**









