ANALYSIS AND PREDICTION OF VECTOR BORNE DISEASE: DENGUE

—

A Report by:

Riya Shah

Krishna Kamath

Simran Bardhan

Rishita Merchant

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

University of Mumbai

2022 – 2023

ANALYSIS AND PREDICTION
OF VECTOR BORNE DISEASES: DENGUE

Submitted in partial fulfillment of the requirements

of the course **Innovative Product Development (IPD) IV**

Year 3, Sem VI Computer Engineering

By

Riya Shah	**60004200118**
Krishna Kamath	**60004200126**
Simran Bardhan	**60004200135**
Rishita Merchant	**60004200159**

Guide:

Prof. Lakshmi Kurup
Assistant Professor

CERTIFICATE

This is to certify that the project entitled “**ANALYSIS AND PREDICTION OF VECTOR BORNE DISEASES: DENGUE**” is a bonafide work of “**Riya Shah, Krishna Kamath, Simran Bardhan, Rishita Merchant**” (**60004200118, 60004200126, 60004200135, 60004200159**) submitted as a project work for the course **Innovative Product Development (IPD) IV, Year 3, Semester VI, TY B.Tech Computer Engineering**

Prof. Lakshmi Kurup

Internal Guide

Dr. Meera Narvekar	**Dr. Hari Vasudevan**
Head of Department	**Principal**

IPD Project Report Approval for BTech Semester VI

This project report entitled **ANALYSIS AND PREDICTION OF VECTOR BORNE DISEASE: DENGUE** by **Riya Shah, Krishna Kamath, Simran Bardhan, Rishita Merchant** is approved for the Innovative Product Development (IPD) IV examination **of Year 3, semester VI, TY B.Tech Computer Engineering**

Examiners

1.--

2.--

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included. We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Riya Shah - 60004200118)

(Krishna Kamath - 60004200126)

(Simran Bardhan - 60004200135)

(Rishita Merchant - 60004200159)

Date:

Abstract

Vector-borne diseases pose a significant threat to global health, as they are transmitted by infectious arthropods such as sandflies, ticks, triatomine insects, and mosquitoes. Among these diseases, dengue fever stands out as one of the most prevalent viral infections worldwide. The warm and humid climates found in tropical regions provide an ideal habitat for the mosquitoes that carry and transmit the dengue virus. Currently, nearly thirty per cent of the global population resides in these regions, making them particularly vulnerable to dengue. However, predicting how climate change will impact dengue infection rates is a complex task. The re-emergence of dengue fever is influenced by a combination of social and environmental factors, including rapid urbanization, increased global travel, and international trade. Additionally, climate change plays a role as rising temperatures can accelerate the reproductive cycle of dengue mosquitoes, leading to more frequent biting and higher transmission rates. To understand and respond effectively to these dynamics, researchers employ advanced analytical techniques, leveraging comprehensive databases and machine learning models to explore the relationship between dengue fever and climate change. These efforts aim to anticipate future changes in dengue transmission patterns and develop strategies to mitigate the impact of climate change on public health.

In the pursuit of a deeper understanding of the link between dengue fever and climate change, temporal statistical analysis plays a crucial role. Researchers rely on meticulously curated databases specifically designed to capture weather variations. These databases provide a wealth of information that can be pre-processed and analyzed using machine learning technologies. By employing various machine learning models, scientists can make accurate predictions and identify potential correlations between dengue infection rates and climate change. This approach enables the development of sophisticated tools and applications that equip healthcare professionals with advanced knowledge and insights into the best healthcare tactics and strategies under specific climatic conditions. Armed with this information, physicians can better prepare and respond to the challenges posed by dengue outbreaks in a changing climate.

The findings from the research on dengue fever and climate change have significant implications for both public health and climate adaptation efforts. As climate change continues to shape our planet, it is essential to understand how it influences the transmission dynamics of vector-borne diseases like dengue. The scientific community has made strides in uncovering the complex interplay between climate variables and dengue infection rates, shedding light on potential future modifications due to climate change. This knowledge enables policymakers and healthcare practitioners to develop targeted interventions, implement effective prevention strategies, and allocate resources more efficiently. By leveraging the power of machine learning and data-driven approaches, the field of dengue research has made substantial progress in its ability to anticipate and respond to the impact of climate change on disease transmission. Ultimately, these advancements contribute to building resilience in the face of a changing climate and safeguarding public health in vulnerable regions.

Table of Contents

List Of Figures

List Of Tables

List Of Abbreviations

Sr. No.	Abbreviation	Expanded form
1	SVR	Support Vector Regression
2	KNN	K-Nearest Neighbours
3	ML	Machine Learning
4	NDVI	Normalized Difference Vegetation Index
5	MSE	Mean Square Error

ANALYSIS AND PREDICTION OF VECTOR BORNE DISEASE: DENGUE

1. INTRODUCTION

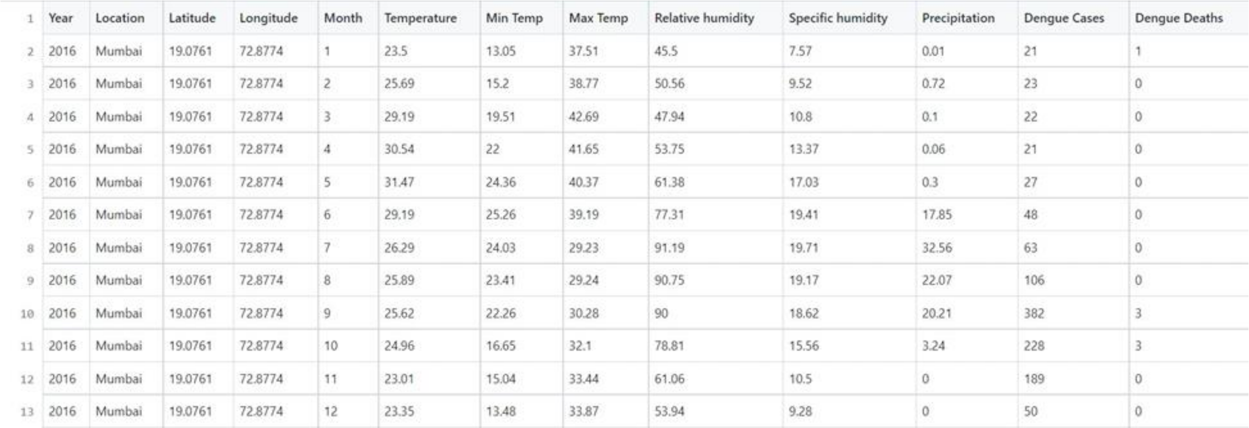
Vector-borne diseases are transmitted by insects like fleas, flies, ticks, and lice, accounting for approximately 17% of all primary infectious diseases. The impact of climate change on the spread of these diseases is well recognized. Dengue virus, ranging from mild asymptomatic illness to severe and potentially fatal conditions like dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS), has witnessed a significant increase in India over the past decade. Factors contributing to this growth include delayed diagnosis and treatment due to limited resources, inadequate mosquito control measures, and the influence of climatic factors such as rainfall, temperature, and moisture content. Understanding the relationship between climate, environment, and dengue incidence remains a critical focus of research.

The analysis and prediction of dengue in the context of climate change have far-reaching implications for public health strategies and policy development. Equipped with this knowledge, healthcare professionals and policymakers can implement targeted interventions and allocate resources effectively to mitigate the impact of dengue outbreaks in a changing climate.

2. SURVEY

2.1. Survey based on dataset

A database is one that manages information in time. According to the problem statement, there are certain attributes that have been included as climatic change parameters. During the data collection period, a basic understanding of which climate parameters were more frequently used was based on published research papers. Research papers were also read for data analysis, which resulted in locating the necessary information and parameters. The next step was to locate a dataset that will be put to use. Because such datasets are very uncommon and not found, hence the conclusion was to create the dataset using the parameters that were discovered to be common in some published papers.



Year	Location	Latitude	Longitude	Month	Temperature	Min Temp	Max Temp	Relative humidity	Specific humidity	Precipitation	Dengue Cases	Dengue Deaths
2016	Mumbai	19.0761	72.8774	1	23.5	13.05	37.51	45.5	7.57	0.01	21	1
2016	Mumbai	19.0761	72.8774	2	25.69	15.2	38.77	50.56	9.52	0.72	23	0
2016	Mumbai	19.0761	72.8774	3	29.19	19.51	42.69	47.94	10.8	0.1	22	0
2016	Mumbai	19.0761	72.8774	4	30.54	22	41.65	53.75	13.37	0.06	21	0
2016	Mumbai	19.0761	72.8774	5	31.47	24.36	40.37	61.38	17.03	0.3	27	0
2016	Mumbai	19.0761	72.8774	6	29.19	25.26	39.19	77.31	19.41	17.85	48	0
2016	Mumbai	19.0761	72.8774	7	26.29	24.03	29.23	91.19	19.71	32.56	63	0
2016	Mumbai	19.0761	72.8774	8	25.89	23.41	29.24	90.75	19.17	22.07	106	0
2016	Mumbai	19.0761	72.8774	9	25.62	22.26	30.28	90	18.62	20.21	382	3
2016	Mumbai	19.0761	72.8774	10	24.96	16.65	32.1	78.81	15.56	3.24	228	3
2016	Mumbai	19.0761	72.8774	11	23.01	15.04	33.44	61.06	10.5	0	189	0
2016	Mumbai	19.0761	72.8774	12	23.35	13.48	33.87	53.94	9.28	0	50	0

Fig. 1 Dataset Created

2.2. Literature Survey

The fundamental requirements of this research were supplemented in a paper [1], which used spatial and temporal data analysis to estimate the influence of climate change on vegetable farming in Tibet. The standardized differential vegetation index (NDVI) was used to collect the data. This paper's main objective was to investigate the impact of climate change on vegetation. In this instance, no forecast for vegetation growth was given. They thus only used spatiotemporal data analysis to identify vegetation changes.

The goal of paper [2] was to comprehend the complexities of seasonal fluctuations in vector-borne diseases. Its methodology was Seasonal and Meteorological Models, which was published on January 10, 2017. Beginning in January 1991, 1993, and 1995, the NNDSS provided dengue, RRV, and BFV case counts and disease incidence rates, respectively.

Paper [3] focuses on how vector-borne diseases are affected by climate change. It gave us some fundamental knowledge of the machine learning techniques employed in climate change prediction. This paper, which was published on February 13, 2010, included a Precis model with a resolution of 50 x 50 km daily temperature and relative humidity for the year 2050. Source: Directorate of National Vector Borne Disease Control Programme (NVBDCP).

The performance of several machine learning algorithms (Linear Regression, Support Vector Regression (SVR), Lasso, and ElasticNet) in the problem of predicting annual global warming from historical measured values is evaluated in paper [4]. There are numerous ideas and projects relating to weather prediction, rainfall prediction, and temperature prediction. A model for forecasting data for the next ten years is trained and tested using linear regression with different input variables such as temperature, carbon dioxide, methane, nitrous oxide, and sulphur hexafluoride. In August of 2020.

This study assesses the various machine learning models that may be employed in paper [5]. Weekly data on climatic conditions and the number of dengue cases were collected for various locations. A dengue search index was created to aid in the development of predictive models in conjunction with climate factors. Several machine learning algorithms, including the support vector regression (SVR) algorithm, step-down linear regression model, gradient boosted regression tree algorithm (GBM), negative binomial regression model (NBM), least absolute shrinkage and selection operator (LASSO) linear regression model, and generalized additive model (GAM), were used as candidate models to predict dengue incidence. The models' performance and goodness of fit were evaluated using the root-mean-square error (RMSE) and R-squared measures. The proposed SVR model achieved superior performance in comparison with other forecasting techniques assessed in this study. The findings can assist the government and community in responding to dengue epidemics as soon as possible. Published on October 16, 2017.

Paper [6] attempts to forecast the likelihood that the occurrence of diseases will enable the populace to be aware of the potential risks and take precautions. It makes use of machine learning and artificial intelligence to determine the methodology that can be used. They use Support Vector Machine classifiers to predict problems with two classes, such as a malaria outbreak, which is either Yes or No. They have thus made a prediction about whether a malaria outbreak will occur or not for a particular range of climatic conditions. Published on February 13, 2021.

In paper [7], this study uses climate variability across six countries in Sub-Saharan Africa over a period of twenty-eight years to propose a machine learning-based model for the classification of malaria incidence. This study, published on January 4, 2021, proposes a framework for classifying malaria incidence based on non-seasonal climate variations using the XGBoost classification model. The USDA Animal Plant Health Inspection Service (APHIS) approach to variable selection and harmonization, identification of variables and data sources, and hypothesis testing were used in this paper.

The paper [8] investigates the relationship between weather predictors, including lagged terms, and dengue incidence in the District of Tawau from 2006 to 2017. Using this information, a forecasting model was created to predict future outbreaks in Tawau. The best fitting model of the methodologies used was the SARIMA with external regressors model, which used maximum temperature, minimum temperature, mean relative humidity, and mean rainfall, as well as a log-likelihood. This study's model demonstrated the ability to forecast potential dengue outbreaks 1 to 4 months in advance.

Studies revealed in the paper [9] that a significant portion of human infections are spread by mosquitoes. As a result, it is critical to investigate the transmission characteristics of mosquito-borne infections. Additionally, it is concluded that mathematical modelling is the ideal tool for analyzing the situation quickly and efficiently in order to comprehend these characteristics. They also went over the fundamentals of mathematical modelling, beginning with its origins. The fundamental SIR model has been thoroughly discussed. Extensive models have also been described and analyzed. Numerical methods for solving these models have also been discussed. This article provides enough basic knowledge about mathematical modelling for mosquito-transmitted viral infections (specifically CHIKV, DENV, ZIKV, and WNV) for researchers who want to get started or are already working in this field. In addition, this survey introduced modellers for emerging viral infections.

There are many effective predictions of the vector borne disease outbreak (Multiclass Classification) of three diseases (Chikungunya, Malaria, Dengue) across the Indian-subcontinent which was stated in paper [10]. The authors tested and refined our model using data collected across India from 2013 to 2017. A Convolutional Neural Network outbreak risk prediction algorithm based on disparate data had been proposed. To the best of our knowledge, no previous works have focused on contrasting data in the field of medical data analysis. Their proposed CNN algorithm has a prediction accuracy of 88%.

Table 1. Literature Survey

SR. No.	AUTHORS, PUBLISHED ON	PURPOSE	METHODOLOGY	DATASET	PROS AND CONS	FUTURE SCOPE
[1]	Guangyong You, M Altaf Arain, Shusen Wang, Shawn McKenzie, Changxin Zou, Zhi Wang, Haldong Li, Bo Liu, Xiaohua Zhang, Yangyang Gu, and Jixi Gao Published 5 September 2019	To find the effect of climate change on vegetable cultivation in Tibet	Spatiotemporal Data Analysis	NDVI Data: Normalized Difference Vegetation Index Climate Data	The main advantage is that they made use of spatiotemporal analysis for their study. The disadvantage of this paper was that they did not make any prediction related to future growth in vegetation.	Hoping to get more data related to asymmetric warming on the regional carbon/water flux and carbon/water balance to enhance their model.
[2]	Margraret D. Stratton, Hanna Y. Ehrlich, Siobhan M. Mor and Elena N. Naumova. Published on 10 January 2017	Gaining insight into the complexities of seasonal variations in vector-borne diseases (VBDs)	Seasonal and Meteorological Models	NNDSS provided case counts and disease incidence rates beginning in January 1991, 1993, and 1995 for dengue, RRV, and BFV, respectively.	Variations in local vector species may contribute to differences in the spatiotemporal patterns of vector-borne diseases. However, the national disease system lacks information on the country of acquisition for key VBDs like BFV and RRV.	Trying to understand the potential underlying factors and characteristics that contribute to the observed seasonal patterns. Predictably, the addition of seasonality and lagged weather data.

[3]	Ramesh C Dhiman, Sharmila Pahwa, G P S Dhillon, Aditya P Dash Published on 13 February 2010	Impact of climate change on vector borne diseases.	Using PRECIS model (driven by HadRM2) at the resolution of 50 x 50 Km daily temperature and relative humidity for the year 2050 was found	Source: Directorate of National Vector Borne Disease Control Programme (NVBDCP)	Efforts were made to evaluate and monitor disease surveillance, supervisor indoor residual spray operations and assessment of therapeutic efficacy and insecticide resistance but couldn't do so.	Hoping that with further strengthening of infrastructure, identification of constraints and policy and better assessment tools available, the threat of climate change on vector-borne diseases in India may be negated.
[4]	Himanshu Vishwakarma Student, Department Of Computer Science & Engineering, IMS Engineering College, Ghaziabad, UP, 201009, India Published in August 2020	This paper evaluates the performance of several Machine Learning algorithms (Linear Regression, Support Vector Regression (SVR), lasso, Elastic Net) in the problem of annual global warming prediction, from previously measured values.	Machine Learning Algorithms Linear Regression Support Vector Regression Lasso Regression	Temperature Carbon dioxide Methane Nitrous oxide Sulphur Hexafluoride	This model predicts only the mean temperature and only 4 greenhouse gas concentrations. It doesn’t explain each week and each country’s data separately. According to the paper, it needs more data for testing and training the model.	Future predictions can be enhanced by incorporating larger datasets and predicting the influence of other factors on temperature rise. Weekly and country specific data can make it more accessible and relevant to a wider audience.

[5]	Pi Guo, Tao Liu, Qin Zhang, Li Wang, Jianpeng Xiao, Qingying Zhang, Ganfeng Luo, Zhihao Li, Jianfeng He, Yonghui Zhang, Wenjun Ma. Published on 16 October 2017	A dengue search index was created to develop predictive models in conjunction with climate factors. The models also incorporated the year and week of observation to account for long-term trends and seasonality.	1. Support Vector Regression (SVR) algorithm 2. Step-down linear regression model 3. Gradient Boosted Regression Tree Algorithm (GBM) 4. Negative binomial regression model (NBM) 5. Least absolute shrinkage and selection operator (LASSO) 6. Linear regression model 7. Generalized additive model (GAM)	2011 - 2014 (1st Jan 2011 - 31st Dec 2014) Weekly dengue cases Baidu search queries Climate factors (mean temp., relative humidity and rainfall)	The main advantage was that it correctly identified that the SVR model can be used in the given scenario. It also gave a comparison chart based on each model. The main disadvantage was that while it did say which model can be used, it did not focus much on predicting future values.	The paper's main aim is to further expand on the models to predict values related to the number of cases with the help of the current dataset.
[6]	Agranee Jha, Sanchit Vartak, Kavya Nair, Anil Hingmire Published on 13 February 2021	Made use of SVM to classify if a malaria outbreak will occur in the given weather conditions.	1. Naive Bayes Method 2. ARIMA 3. Artificial Neural Network 4. Support Vector Machine (SVM)	Malaria data - National Vector Borne Disease Control Program Climate Data - data.gov.in For data not found 'Replace missing Value' tool of Weka has been used	The main advantage is that it provides the results of each model and suggests that SVM model be used in the given scenario. The main disadvantage of the paper is that it does not predict the values for the future and it does not give further details for the other models they have tested.	In the future, more data that is further localized can be used to get more accurate predictions. Also, such models can be scaled up to include the country and can be used on different other diseases.

[7]	Odu Nkiruka, Rajesh Prasad, Onime Clement Published on 10 January 2021	The purpose of a machine learning based model classification of malaria incidence using climate variability across six countries of Sub-Saharan Africa over a period of twenty-eight years.	This research paper proposes a framework that classifies malaria incidences based on non-seasonal climate variations using the XGBoost classification model.	USDA APHIS use systematic variable selection, data source identification and hypothesis testing for harmonization and analysis.	This study develops a machine learning model for binary prediction of malaria outbreaks using non-seasonal changes in climatic factors. Time-series models such as ARIMA and SARIMA are utilized.	Future work would involve obtaining an adequate dataset for confirmed malaria incidence, possibly as time-series data that can seasonally stratify important malaria seasons to enhance the real-time prediction by the system.
[8]	Vivek Jason Jayaraj, Richard Avoi, Navindran Gopalakrishnan, Dhesi Baha Raja, Yusri Umasa Published on 8 June 2019	This study analyzes the relationship between weather predictors including its lagged terms, and dengue incidence in the District of Tawau over a period of 12 years, from 2006 - 17. A forecasting model proposed to predict future outbreaks in Tawau was developed.	These variables were then employed in 3 different methods: a multivariate Poisson regression model, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model and a SARIMA with external regressors for selection.	Monthly dengue incidence data, mean temp., max temp., min temp., mean relative humidity mean rainfall over a period of 12 years from 2006 to 2017 in Tawau - Tawau District Health Office and the Malaysian Meteorological Department	A forecasting model was developed to predict future outbreaks in Tawau. Three different methods were employed using various variables, with SARIMA with external regressors emerging as the best fitting model among the methodologies utilized.	The findings align with previous studies conducted in various countries, highlighting the significant predictive power of temperature and humidity on dengue incidence. However, further progress in the model necessitates larger and more diverse datasets.

[9]	Ramakant Prasad Surendra Kumar Sagar Shama Parveen Ravins Dohare Published on 19 August 2022	This study will be helpful for mathematical modellers in vector-borne diseases and ready-made material in the review for future advancement in the subject.	Based on different mathematical models like SEIRS, SEIR, SIRS, SIR, SEIS, SEI, SIS, SI and the Basic SIR model	Weekly dengue cases Climate factors (mean temp., relative humidity and rainfall) used a small-world network to explain the dynamics of a simple disease model.	This study will be helpful for mathematical modelers in vector-borne diseases and ready-made material in the review for future advancement in the subject. This study will not only benefit VBDs but also will enable ideas for other illnesses.	Here, mathematical modeling is a perfect tool to deal with this problem. Although it is very challenging to convert the real problem into a mathematical model in an ideal way, it may be constructed under certain constraints.
[10]	Sandali Raizada, Shuchi Mala, Achyut Shankar Published on 8 December 2020	Finding a relation between environmental factors and outbreak of the disease and supported a model that used classification algorithms.	We seek advice and suggestions from experts to extract meaningful characteristics. We predict the severity of the outbreak in various regions of the country. We also use statistical analysis and mainly to propose a CNN multimodal disease outbreak prediction (CNN-MDOP) algorithm.	Meteorological Data - The data provided by the weather department include temperature, humidity and rainfall over a span of five years from the 13-17 Demographical Data -The data included the number of positive cases in the country arranged state-wise for 13-17.	They seeked advice and suggestions from experts to extract meaningful characteristics. We predict the severity of the outbreak in various regions of the country. We also use statistical analysis and mainly to propose a CNN multimodal disease outbreak prediction (CNN-MDOP) algorithm.	Limited search has been done with considering the years from 13 to 17. And already a lot of study around this subject has already been done.

2.3. Outcome of survey

The survey helped us understand how we can create our own dataset and which machine learning models are prominently used in analysis of vector borne diseases. In paper [1], we discovered how climate change affects various nations and the machine learning (ML) models that were employed to predict climate change in Tibet, Southwest China. We received in-depth information about seasonal variations in the spread of vector-borne diseases in Australia in Paper [2]. In papers [3] and [4], analysis of climate change has been done. It is also seen how these climate changes affect vector bourne diseases. The remaining papers primarily focus on the various machine learning models that they have used for prediction of vector bourne diseases (mostly malaria) and have done a comparitive study of each model.

3. NEED OF THE PRODUCT

3.1. Drawbacks of the existing system

Numerous studies have explored the impact of climate change on vector-borne diseases using machine learning models like linear regression. However, these studies fall short in providing predictions for the future spread of such diseases in response to climate change. Our research addresses this gap by introducing a forecasting tool that utilizes machine learning techniques. Additionally, we present a decision support system that categorizes dengue incidence into high and low target classes by leveraging climate variability. Our paper contributes to the field by offering a more comprehensive approach to understanding and predicting the implications of climate change on the spread of vector-borne diseases.

3.2. Applications of the product

Our product has immense potential in enhancing medical research and improving the understanding of vector-borne diseases, particularly in the context of climate change. It plays a crucial role in planning and preparedness for disease outbreaks triggered by specific climatic conditions. With its advanced ability to categorize and analyze vast amounts of medical data, our product becomes an effective tool in healthcare. By leveraging machine learning techniques, it empowers researchers to forecast future trends and make well-informed decisions in a timely manner. Its capacity to handle large datasets and provide predictive insights holds tremendous promise in advancing healthcare practices and decision-making. This invaluable tool enables healthcare professionals to delve deeper into medical data, identify patterns, and make informed decisions to enhance patient outcomes and tackle emerging challenges in the field.

4. PROBLEM FORMULATION

4.1. Problem Formulation

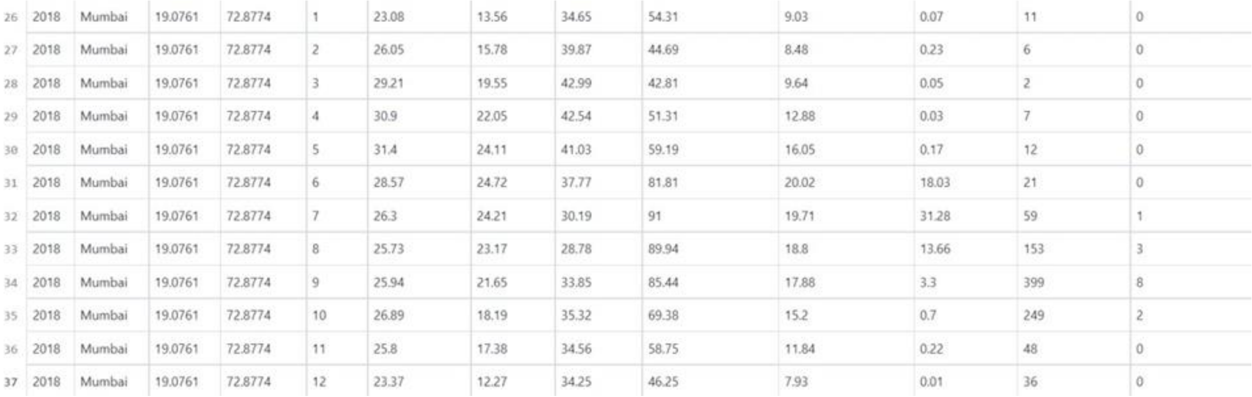
In our research, we faced challenges due to the limited availability of comprehensive temperature and rainfall datasets covering all months and years. To overcome this, we proactively gathered relevant data by extensively searching various websites. Another hurdle was obtaining information on the monthly incidence of a specific vector-borne disease. Consequently, we focused on acquiring data specifically related to dengue cases in Mumbai. Additionally, we conducted a comparative analysis of various machine learning models to assess their predictive capabilities using the available data. ***Owing to the fact that the dataset related to climate data and dengue cases is not present, we have created our own dataset.***

4.2. Product Objectives

- The dataset created will have data related to climate and dengue for different location of India.
- Our product utilizes gathered data to predict the impact of climate change on vector-borne diseases. Before making predictions, we carefully evaluate multiple machine learning algorithms to ensure accurate and reliable results.
- Develop a suitable model for dengue outbreak prediction by comparing various models, experimenting with the data we have, and taking into account the accuracy we receive.

4.3. Novelty

There are no climate and dengue case datasets available online for various locations in India. Thus, we have constructed our own dataset consisting of various parameters, such as the location's geometrical coordinates, climate data including temperature (mean, minimum, and maximum), relative and specific humidity, precipitation, and dengue cases for a given month and year. This dataset was compiled using a variety of online resources and news articles pertaining to climate or dengue.



26	2018	Mumbai	19.0761	72.8774	1	23.08	13.56	34.65	54.31	9.03	0.07	11	0
27	2018	Mumbai	19.0761	72.8774	2	26.05	15.78	39.87	44.69	8.48	0.23	6	0
28	2018	Mumbai	19.0761	72.8774	3	29.21	19.55	42.99	42.81	9.64	0.05	2	0
29	2018	Mumbai	19.0761	72.8774	4	30.9	22.05	42.54	51.31	12.88	0.03	7	0
30	2018	Mumbai	19.0761	72.8774	5	31.4	24.11	41.03	59.19	16.05	0.17	12	0
31	2018	Mumbai	19.0761	72.8774	6	28.57	24.72	37.77	81.81	20.02	18.03	21	0
32	2018	Mumbai	19.0761	72.8774	7	26.3	24.21	30.19	91	19.71	31.28	59	1
33	2018	Mumbai	19.0761	72.8774	8	25.73	23.17	28.78	89.94	18.8	13.66	153	3
34	2018	Mumbai	19.0761	72.8774	9	25.94	21.65	33.85	85.44	17.88	3.3	399	8
35	2018	Mumbai	19.0761	72.8774	10	26.89	18.19	35.32	69.38	15.2	0.7	249	2
36	2018	Mumbai	19.0761	72.8774	11	25.8	17.38	34.56	58.75	11.84	0.22	48	0
37	2018	Mumbai	19.0761	72.8774	12	23.37	12.27	34.25	46.25	7.93	0.01	36	0

Fig 2. Screen capture of Dataset

4.4. Scope of the project

We intend to publish the datasets we've compiled. It would assist them in thoughtfully preparing for any disease outbreak induced by specific climatic conditions. Along with machine learning algorithms, this network may be one of the most effective healthcare instruments for classifying and analysing vast amounts of medical data and predicting future trends. Due to the inability to acquire evaluation or verification of the dataset from a domain expert, the project's scope is restricted to the local domain, despite the product's applicability in numerous fields.

5. PROPOSED DESIGN

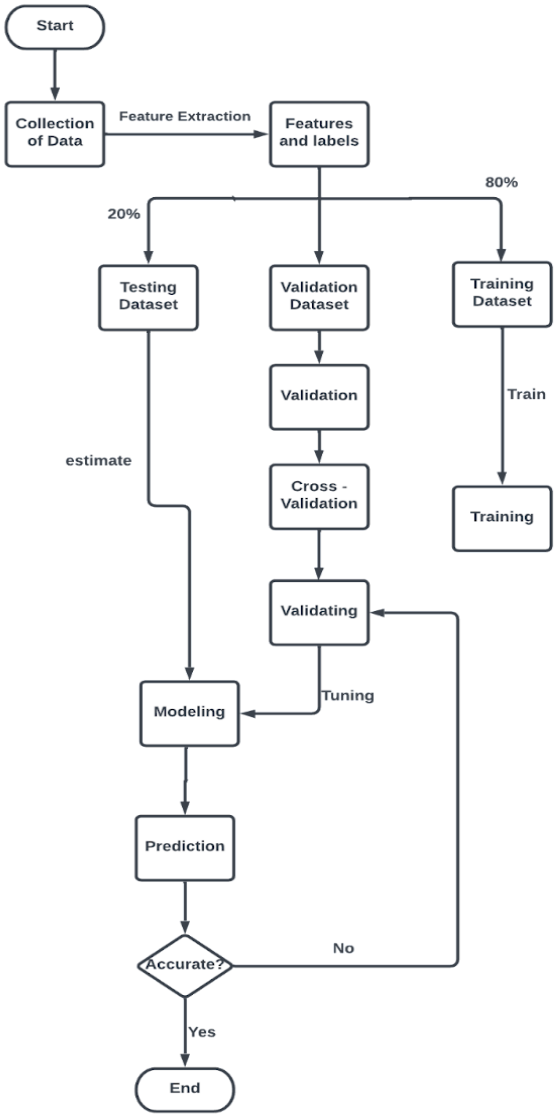


Fig. 3. Proposed Design

6. IMPLEMENTATION

6.1. Dataset

The dataset consists of 85 entries that indicate the monthly occurrence of dengue. The dataset consists of the following attributes -

Table 2. Attributes and Parameters

ATTRIBUTES	EXPLANATION
Year	Year in which required data was recorded
Location	Location of the place where the climatic data and dengue data is collected
Latitude	Geometric latitude of the location
Longitude	Geometric longitude of the location
Month	Month in which required data was recorded
Temperature	Location's typical temperature for the specified month in the specified year
Min Temp	Location's lowest temperature during the specified month in the specified year
Max Temp	The location's highest temperature for the specified month in the specified year
Relative humidity	Location's relative humidity for the specified month in the specified year
Specific humidity	Location-specific humidity for the given month in the given year
Precipitation	cms of rainfall recorded for the location
Dengue Cases	Cases of dengue reported in the given month and year
Dengue Death	Number of deaths reported in the given month and year

6.2. Modules implementation

The incidence of vector-borne diseases, particularly Dengue, has been predicted in this study using a variety of machine learning models based on climate variations in India. A dataset of historical weather and disease data was used to train the models. The models can be utilized for forecasting the prevalence of vector-borne diseases by taking into account the expected alterations in weather patterns.

The model's validity was confirmed using cross-validation methods on the dataset. Specifically, K-fold cross-validation was utilized in this study, which involves dividing the input dataset into K groups or folds of equal size. In each learning iteration, the prediction function trains on K-1 folds and tests on the remaining fold to evaluate the model's performance.

Linear regression, Support vector regression, Decision Tree Regressor, K-means clustering, K-Nearest Neighbours (KNN), Lasso regression, Ridge regression, and Nu Support vector regression are the machine learning models that have been applied in this study.

Linear Regression:

A supervised learning approach used for regression tasks is linear regression. It deduces that the input features and the response attribute have a linear relationship. The line that minimizes the sum of squared discrepancies between the expected and actual values is the best fit line. Although it is easy to understand and simple, it might struggle with complex relationships.

Support Vector Regression:

The regression variant of Support Vector Machines (SVM) is called Support Vector Regression (SVR). It locates a hyperplane that maximizes the margin while permitting a predetermined amount of faults using support vectors. The goal of SVR is to reduce the epsilon-insensitive loss, which is the loss that ignores errors within a given margin (epsilon). By utilizing kernel functions, it can manage nonlinear interactions. When working with small to medium-sized datasets, SVR is effective.

Decision Tree Regressor:

To make predictions, the Decision Tree Regressor constructs a binary tree structure. It divides the data into branches and leaf nodes depending on the feature values. The target variable is predicted by each leaf node. It can handle categorical and numerical data and captures non-linear correlations. The Decision Tree Regressor carries a risk of overfitting and may not apply well to newly collected data.

K-means Clustering:

K-means Clustering is a type of unsupervised learning method. Data is divided into k clusters, each of which will include the data point that corresponds to its nearest mean. It reduces the total squared distance between the cluster centroids and the data points. In advance, the number of clusters (k) must be specified. K-means is effective, however it might converge to different conclusions depending on the original cluster allocations.

KNN(K-Nearest Neighbor):
KNN is a supervised learning technique applied to both regression and classification. The k nearest neighbors in the training set are taken into account for predicting the target variable.For classification and regression, it assigns the majority class or averages the target values of the k nearest neighbors. KNN can handle non-linear relationships and does not require explicit model training. It is sensitive to the choice of distance measure and, for large datasets, can be computationally expensive.

Lasso Regression:
A regularization term is part of the Lasso Regression linear regression model. By incorporating the absolute values of the coefficients into the loss function, it conducts both feature selection and regularization. The less significant features are shrunk towards zero, which promotes sparse solutions. Lasso Regression can be helpful for handling multicollinearity and feature selection. When working with high-dimensional datasets, it is efficient.

Ridge Regression:
A regularization term is part of the Ridge Regression linear regression model. The coefficients' squared values are added to the loss function. Overfitting is avoided, and the influence of highly correlated features is diminished. Multicollinearity can be handled via ridge regression, which also offers more reliable solutions. It is appropriate when the dataset has a large number of correlated features.

Nu Support Vector Regression:
A variant of Support Vector Regression (SVR) is called Nu Support Vector Regression(NuSVR). It adds a new parameter (nu) that regulates the margin width and the number of support vectors. The goal of Nu SVR is to determine the ideal balance between model complexity and accuracy. Through the kernel functions, it can handle non-linear relationships. When there is uncertainty on the optimum amount of support vectors, Nu SVR is helpful.

7. EXPERIMENTATION & RESULT

Table 3. Comparative study of the implemented models.

Model	*Accuracy and Mean Square Error*	*STRENGTHS*	*WEAKNESSES*
Nu Support Vector Regression	MSE - 2860.05	Worked better with the weather data than SVR as it is a multi-step ahead prediction and optimizes the values.	The model is inefficient in predicting the fine patterns in the data
Linear Regression	MSE - 3024.64	Worked best with the data that had a linear trend.	The model is not fit for forecasting time-series data
Support Vector Regression	MSE - 3163.30	Worked well with collected weather data.	The model is not fit for forecasting non-seasonal data
Decision Tree Regressor	MSE - 3076.71	It works well with our dataset's non-linear connection data.	It is overfitting the tree hence resulting in some inaccurate results.
K-Means Clustering	-	K-Means returned clusters which were easily interpretable and visualizable.	Clustering data of varying sizes and densities was difficult to cluster.
KNN	-	No need for a training phase: Adding new data is straightforward. Simple to use.	With big datasets and plenty of dimensions, it struggles. able to detect erratic data, missing values, and outliers.
Lasso Regression	MSE - 3049.40	Feature Selective, it identifies the most important predictors and will exclude the less important ones, which helps in	Lasso Regression does not perform well on non-linear data as it assumes the relationship between the attributes and

		accurate and reliable predictions. Thus, it will help recognize the attribute which has the strongest impact on dengue outbreak and help in developing strategies to prevent or control dengue outbreak.	response variable is linear.
Ridge Regression	MSE - 3355.08	Similar to lasso regression, it helps in identifying the attribute which highly affects the response variable to help in preventing or controlling the variable. It also reduces the risk of overfitting of models due to the regularization term.	Ridge regression is sensitive to the regularization parameter being used. Thus, it may result in overfitting or underfitting of data. Also, it assumes the relationship between the attributes and response variable (dengue cases) to be linear, but the relation is actually non-linear.

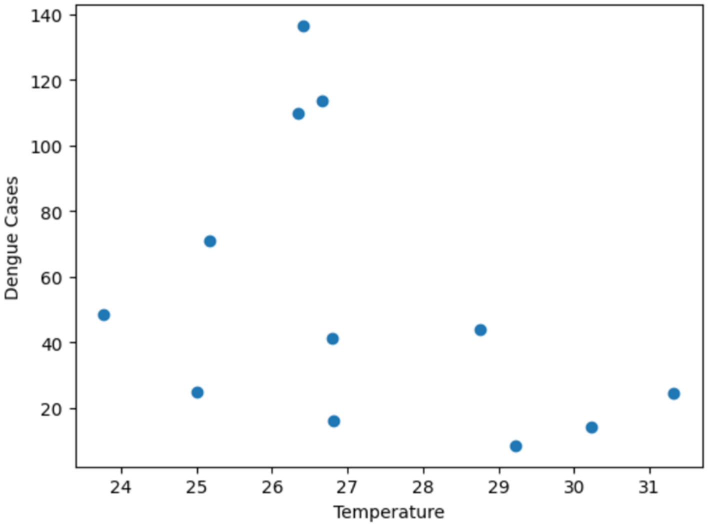


Fig. 4. Prediction of dengue cases with respect to change in temperature (2023)

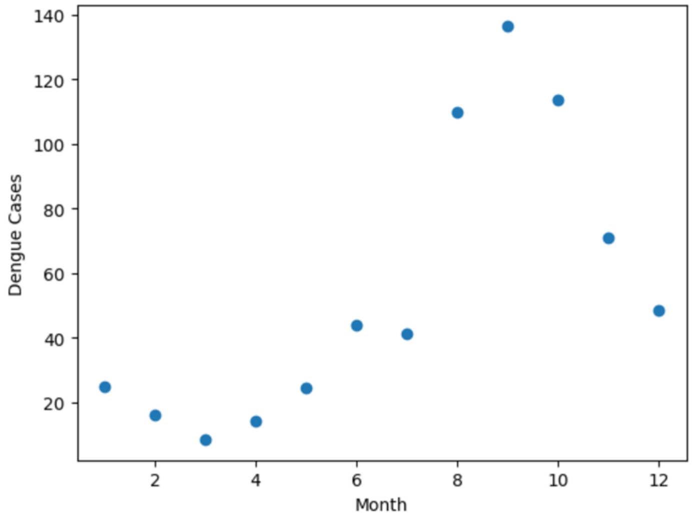


Fig. 5. Prediction of dengue cases monthwise (2023)

The percentage of cases that a machine learning model correctly categorizes is known as accuracy. It is calculated by dividing the total number of accurate forecasts by the sum of all guesses across all classes.

The accuracy is calculated using the formula below: Total Predictions / (Number of Correct Predictions).

The model's predictions' mean squared error is quantified by the Mean Squared Error (MSE). The squaring strategy is widely utilized in regression situations since it penalizes greater errors more severely. A lower MSE value indicates better performance and signifies that the projected values are generally closer to the actual ones.

As noted in Table 3, NuSVR has the least MSE of all the models mentioned above. Therefore, after performing K- Fold Cross Validation with 10 splits and an MSE of 78.9, the NuSVR model is chosen for case prediction.

8. CONCLUSION

In conclusion, this project presents a comprehensive analysis of the application of machine learning techniques in understanding and predicting the incidence of vector-borne diseases, specifically focusing on the impact of climate variability. The developed model, utilizing the NuSVR algorithm, exhibits a high level of accuracy in forecasting dengue cases, as evidenced by its low Mean Squared Error (MSE). The utilization of K-Fold Cross Validation further strengthens the reliability and generalizability of the model's predictions.

The findings of this project have significant implications for decision-making in the prevention and control of future dengue outbreaks. By uncovering the intricate relationship between climatic factors and vector-borne diseases, policymakers, public health officials, and healthcare practitioners can make informed decisions and implement proactive measures to mitigate the impact of dengue on affected populations.

However, further investigations are warranted to improve the predictive potential of the models presented in this study. This entails the accumulation of a more extensive and confirmed dataset for dengue illness, preferably with resolutions similar to or finer than the climatic observations utilized during model training. Such efforts will enhance the accuracy and robustness of the predictive models, enabling more precise forecasts and proactive planning for dengue outbreak prevention.

9. REFERENCES/BIBLIOGRAPHY

[1] M. A. A. S. W. S. M. C. Z. Z. W. H. L. B. L. X. Z. Y. G. a. J. G. Guangyong You, "The spatial-temporal distributions of controlling factors on vegetation growth in Tibet Autonomous Region, Southwestern China," 2019.

[2] H. Y. E. S. M. M. a. E. N. N. Margraret D. Stratton, "A Comparative analysis of three VBDs across Australia using seasonal and meteorological models," 2017.

[3] S. P. G. P. S. D. A. P. D. Ramesh C Dhiman, "Climate change and threat of vector-borne diseases in India.," 2010.

[4] D. O. C. S. &. E. I. E. C. G. U. 2. I. Himanshu Vishwakarma Student, "Climate Change Analysis Using Machine Learning," 2020.

[5] Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma W. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017 Oct 16;11(10):e0005973. doi: 10.1371/journal.pntd.0005973. PMID: 29036169; PMCID: PMC5658193.

[6] Agranee Jha, Sanchit Vartak, Kavya Nair, Anil Hingmire, 2021, Malaria Outbreak Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03),

[7] Odu Nkiruka, Rajesh Prasad, Onime Clement, Prediction of malaria incidence using climate variability and machine learning, Informatics in Medicine Unlocked, Volume 22, 2021

[8] Nkiruka, Odu & Prasad, Dr & Onime, Clement. (2021). Prediction of Malaria Incidence using Climate Variability and Machine Learning. Informatics in Medicine Unlocked. 22. 10.1016/j.imu.2020.100508.

[9] Prasad, R., Sagar, S.K., Parveen, S. et al. Mathematical modeling in perspective of vector-borne viral infections: a review. Beni-Suef Univ J Basic Appl Sci 11, 102 (2022). https://doi.org/10.1186/s43088-022-00282-4

[10] 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. APPENDIX