Smart Health: A Machine Learning Approach to Predict Vector-Borne Diseases

[1]G Saiteja, [2]C Harini, [3]G. Rithika, [4]Manan Rasool
[1,2,3]Student,[4]Assistant Professor
[1,2,3,4,]Department of Computer Science and Engineering
[1,2,3,4,]Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

ABSTRACT :

Vector-borne diseases, like malaria, dengue, and chikungunya, are spread by insects like mosquitoes and ticks, posing significant health risks, particularly in tropical and subtropical regions. These diseases are especially prevalent in areas with limited access to healthcare, where they place a considerable burden on public health, with India experiencing millions of cases each year. On a global scale, these diseases account for over 17% of infections caused by pathogens such as bacteria, viruses, and parasites. Malaria, in particular, is responsible for around 219 million cases and more than 400,000 deaths annually. Detecting and controlling these diseases early is difficult because of factors like unpredictable weather, environmental changes, differences in regions, and varying socio-economic conditions. This research aims to address these complexities by analyzing historical data to predict the spread of vector-borne diseases in India. By employing advanced machine learning techniques, such as Random Forest and Logistic Regression, we have developed models that accurately forecast potential outbreaks. These models are crucial in identifying high-risk areas, enabling healthcare professionals to respond promptly and implement targeted interventions. By improving prediction accuracy, the study aims to enhance public health responses, optimize disease control measures, and contribute to more effective strategies for prevention.

1. INTRODUCTION:

Vector-borne diseases (VBDs) are illnesses caused by germs that spread to humans through bites from insects like mosquitoes and ticks. Common diseases include malaria, dengue, chikungunya, and Kala azar. These diseases are seasonal and often spread due to changes in the environment. Vector-borne diseases are a major global health issue, affecting millions of people worldwide. They are especially troublesome in developing countries like India, Bangladesh, and Nepal Many people lose their lives because of late diagnosis, lack of recognition, and delayed treatment of these diseases. Dengue, for example, is widespread in 129 countries, causing 96 million cases and around 40,000 deaths annually. In rural areas, controlling these diseases has become even more challenging due to a lack of proper healthcare facilities and infrastructure. Therefore, it is crucial to find solutions to control these outbreaks using modern technologies such as Information Technology and Artificial Intelligence (AI). There is a growing need for tech-based platforms that can bridge the gap between urban and rural areas by providing access to healthcare through AI. With the advancement of AI, intelligent machines have emerged that can monitor and report health data and patient conditions continuously in various medical settings. By utilizing AI, we aim to equip healthcare workers with tools to identify high-risk areas and take timely action to manage these diseases.

1.1 OBJECTIVE:

The primary goal of this study is to create a predictive model using machine learning techniques to forecast the occurrence of vector-borne diseases in India. This will involve identifying key factors related to these diseases and developing a model that utilizes methods like Random Forest and decision trees, which have proven to be highly accurate in predicting potential outbreaks. Additionally, the study will focus on testing and validating the effectiveness of the proposed model to ensure its reliability and accuracy in real-world applications.

1.2 LITERATURE SURVEY:

Title: Vector Borne Disease Prediction using Multilayer Perceptron - A Machine Learning Approach

Authors: V. Janani, N. Maadhuryaa, D. Pavithra, S. Ramya Sree
Year: 2020
Description:
Machine learning (ML), a subset of Artificial Intelligence (AI), enables systems to learn from data and experience without needing human guidance. The key objective of ML is to recognize trends and patterns, continually improving with each iteration. Vector-borne diseases, especially those caused by mosquitoes, have become increasingly common. This study applies the Multilayer Perceptron (MLP) algorithm to accurately predict these diseases. The method uses an optimization approach with multiple stages. The prediction process includes: 1) Data preprocessing with the Sequential Minimal Optimization (SMO) algorithm, which resolves quadratic programming issues in Support Vector Machines (SVM); 2) Feature selection through MLP combined with backward logistic regression for risk analysis; and 3) Feature reduction for dimensionality reduction in the dataset. The WEKA tool is used for implementation, which includes machine learning algorithms for tasks like classification, regression, clustering, and visualization.

Title: Prediction of Zika Virus Using a Multilayer Perceptron Neural Network (MLPNN) on Cloud B

Authors: Mahalakshmi, G. Suseendran
Year: 2020
Description:
The Zika virus, which spread from Brazil to the Americas, poses a significant public health threat. Without reliable forecasts on the virus's spread, public health agencies struggled to allocate resources efficiently. This paper proposes using a hybrid optimization approach combining a multilayer perceptron with a probabilistic optimization technique to improve prediction accuracy. The MATLAB program integrates machine learning and AI techniques, significantly reducing forecast time while maintaining high accuracy. The results are encrypted with Advanced Encryption Standard (AES) and Triple Data Encryption Standard (TDES), ensuring secure communication of patient information. The approach demonstrates a 91.25% accuracy with minimal processing time.

Title: Artificial Intelligence Model for Early Prediction of Cardiac Functionality Using Multilayer Perceptron

Authors: Sivaranjani R., Dr. N. Yuvaraj
Year: 2021
Description
Cardiovascular diseases (CVDs) are a leading cause of death worldwide, with significant disparities in premature mortality rates between high- and low-income countries. Early prediction plays a crucial role in reducing such deaths. This paper presents an AI-based system for predicting cardiac conditions. The system follows three key steps: 1) Collecting a dataset containing 13 clinical attributes, 2) Training the dataset with an artificial neural network using a backpropagation technique, and 3) Developing a user interface that allows practitioners to input clinical data and receive predictions on cardiac health. This system is designed to assist healthcare professionals in providing early diagnoses and reducing healthcare costs.

Title: Prediction of Disease Level Using Multilayer Perceptron Artificial Neural Network for Patient Monitoring

Authors: Md. Osman Goni Nayeem, Maung Ning Wan, Md. Kamrul Hasan
Year: 2020
Description:
Artificial Neural Networks (ANNs) have proven effective for early disease diagnosis. This study utilizes ANN to predict the onset of three diseases: heart disease, liver disorder, and lung cancer. Using a feed-forward backpropagation neural network with multilayer perceptron, the model classifies liver disorder into four categories: normal, early abnormal, abnormal, and severe. The system demonstrates reliable performance with minimal errors in predicting disease stages, showing its potential for aiding medical diagnoses.

1.3 EXISTING SYSYTEM:

This paper introduces an innovative approach for classifying Vector Borne Disease fevers using online learning methods. The proposed approach is highly practical, as it can learn from only a few training samples. Over time, it can continuously update its learning from new data without the need for retraining or redeploying the model. Additionally, the paper provides an in-depth evaluation of different machine learning techniques for predicting Vector Borne Disease fevers.

The proposed system takes recorded patient symptoms and diagnostic results as input. Offline classifiers are used to generate baseline scores, ensuring the chosen features are suitable for classifying Vector Borne Diseases. The main advantage of the online detection model is its ability to accurately identify patients at high risk for Vector.

1.3.1 EXISTING SYSTEM ADVANTAGE:

- Works well with small datasets and minimal training data.
- Can continuously learn from new data without the need for model retraining.
- Offers a probabilistic framework for class predictions, aiding in more informed decision-making.

1.4 PROPOSED SYSTEM

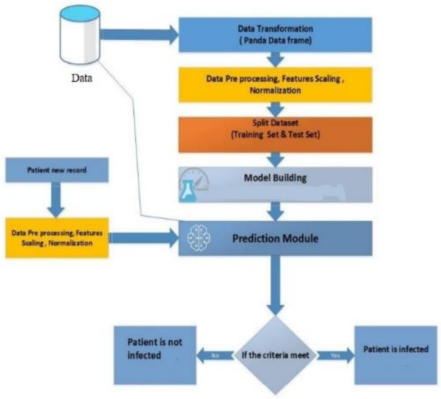
The World Health Organization (WHO) reports that vector-borne diseases are responsible for more than 17% of all infectious diseases globally. Malaria transmission occurs in 97 countries, putting approximately 3.4 billion people at risk. Over 40% of the global population is exposed to vector-borne diseases. In 2006, Chikungunya outbreaks were reported in several countries, including India, and by 2013, the United States also recorded cases of vector-borne diseases. The worldwide vector control efforts identify vector control as a critical strategy for preventing and responding to such outbreaks.

This paper focuses on the prediction of outbreaks of three vector-borne diseases—Chikungunya, Malaria, and other Vector-Borne Diseases—across the Indian subcontinent. Our model, refined through data collected across India, uses a Logistic Regression algorithm to predict.

1.4.1 PROPOSED SYSTEM ADVANTAGE

- It is straightforward to implement, easy to understand, and efficient to train
- It can be easily adapted for multiple classes (through multinomial regression) and offers a natural probabilistic approach to predicting classes
- Additionally, it is highly efficient in classifying new, unknown data.

1.5 SYSTEM ARCHITECTURE:



EXPLAINATION:

The architecture details a machine learning pipeline for disease prediction. Initially, the data undergoes preparation, involving cleaning and transformation for analysis suitability. Subsequently, the prepared data is partitioned into two sets: a training set for model training and a testing set for performance evaluation. The model is then trained on the training data, learning patterns and relationships to make accurate predictions. Once trained, the model is deployed to predict the infection status of new patients based on their input features.

2. DESCRIPTION:

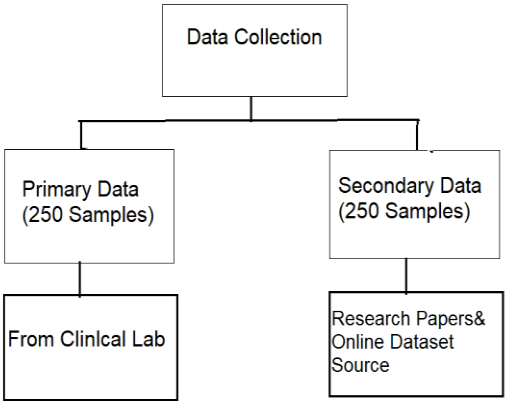
2.1 GENERAL:

Preventing and controlling vector-borne disease outbreaks is a significant challenge worldwide, especially in India, due to its diverse geography and environmental conditions. Diseases like Malaria, Chikungunya, Zika, and Yellow Fever are spread by infected mosquitoes, and every year, many lives are lost due to delayed detection and treatment. While several models have been developed for specific vector-borne diseases, a unified prediction model for all such diseases is still missing. This highlights the need for a machine learning-based model to predict and manage these diseases more effectively in India.

2.2 METHODOLOGIES:

Data collection

Primary data was collected from 250 patients' Vector Borne Disease test reports, including Ns1 and IgM results, at Tuli Diagnostic Centre in Amritsar. The symptom details for each patient were obtained via telephone



Raw data preparations

The foundation for building any predictive model is a high-quality, error-free dataset. Both the quantity and quality of the data significantly impact the model's performance. Preparing the data for analysis by organizing unstructured and semi-structured data into a structured format is crucial to ensure optimal model training and prediction.

Feature selection

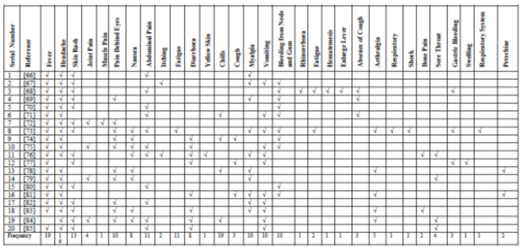
Feature selection plays a critical role in developing an effective machine learning prediction model. A well-organized dataset is essential for building a successful model. Therefore, it is important to identify the relevant features associated with Vector Borne Diseases. In this section, the researcher analyzed twenty studies on the symptoms and parameters of these diseases. Based on this analysis, key symptoms were selected as features. A table was created listing these symptoms along with their occurrence frequency. The table emphasizes the symptoms that are most common and considered significant for early detection of Vector Borne Diseases using machine learning methods.

Identification of features pertaining to Vector Borne Disease

This section aims to identify the symptoms and parameters necessary for diagnosing Vector Borne Diseases based on a review of existing literature. To do so, the researcher examined twenty research papers related to various symptoms of Vector Borne Diseases, which are summarized in Table 4.2.

Table 4.2 presents the key factors for detecting Vector Borne Diseases, along with the frequency of each symptom. Figure 4.3 visually illustrates these important factors, such as fever, muscle pain, and vomiting, using bar graphs. The researcher also ranks each symptom, as detailed in Table 4.2.

Table 4.2: Identification of features pertaining to Dengue Disease



Serial Number	Reference	Fever	Headache	Skin Rash	Joint Pain	Muscle Pain	Pain Behind Eyes	Nausea	Abdominal Pain	Itching	Fatigue	Diarrhoea	Yellow Skin	Chills	Cough	Myalgia	Vomiting	Bleeding from Node and Gum	Rhinorrhoea	Fatigue	Hematemesis	Enlarge Lever	Absence of Cough	Arthralgia	Respiratory	Shock	Bone Pain	Sore Throat	Gastric Bleeding	Swelling	Respiratory System	Petechiae
1	[66]	√	√	√					√							√																
2	[67]	√	√	√						√						√	√	√														
3	[68]	√	√	√					√									√	√	√	√	√	√						√			
4	[69]	√	√	√			√									√		√					√									
5	[70]	√	√	√					√									√														
6	[71]	√	√						√					√			√	√					√									
7	[72]	√	√	√	√	√	√																									
8	[73]	√	√	√			√	√	√		√					√	√	√		√				√	√	√			√		√	
9	[74]	√	√				√	√				√		√	√			√														
10	[75]	√	√		√		√	√	√			√					√	√														
11	[76]	√	√	√				√	√	√		√	√			√	√										√	√				
12	[77]	√		√								√			√		√												√	√		
13	[78]	√	√				√	√						√		√								√								√
14	[79]	√	√		√		√	√								√												√				
15	[80]	√	√	√					√									√														
16	[81]	√	√									√			√	√	√	√						√								√
17	[82]	√	√	√			√		√							√	√															
18	[83]	√	√	√			√	√				√				√	√							√			√					
19	[84]		√	√	√		√	√	√			√		√			√							√				√				
20	[85]	√	√	√					√			√					√											√				
Frequency		19	19	13	4	1	10	8	11	2	11	8	1	19	3	10	10	10	1	2	1	1	3	5	1	1	2	4	3	1	1	2

Split dataset module

This section outlines the process of dividing a large dataset into training and test sets for creating a prediction model for vector-borne diseases. The dataset was split into two parts with a 70:30 ratio, where the first portion is used for training and the second for testing, as illustrated in Figure 5.6

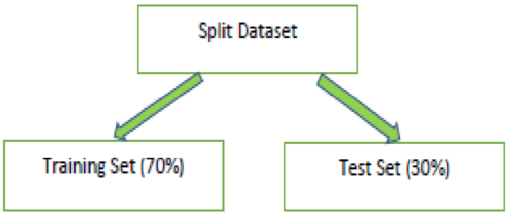


Fig. 5.6 Ratio of data split in 70: 30 ratios of dataset

Model building module

This section details the use of the Logistic Regression Classification algorithm in building the proposed prediction model, along with a comprehensive explanation of its implementation.

Prediction model

This section explains the cutting-edge techniques employed in the proposed prediction model for diagnosing Vector Borne Diseases. The model is capable of handling two scenarios: symptoms without clinical reports and symptoms with clinical reports, as shown in Figure 5.9.

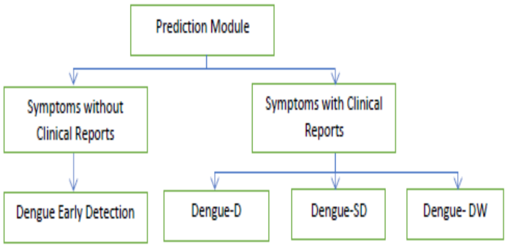


Fig 5.9: Prediction module of proposed prediction model for vector borne diseases (Dengue)

2.3 TECHNIQUE USED:

Logistic Regression & Random Forest:
Random Forest is a collection of many Decision Trees working together. Instead of relying on a single Decision Tree, Random Forest combines the results of multiple trees to make better predictions. This collective approach enhances accuracy by reducing errors through averaging the outputs from various trees.

Random Forest is particularly useful in predicting disease outbreaks because it considers a wide range of factors and identifies patterns that a single tree might miss . Logistic Regression is a predictive modeling technique that learns from labeled examples (data with known outcomes). In simple terms, it predicts an outcome (e.g., whether someone is sick) based on input information (e.g., symptoms) when the outcomes are categorical (grouped).

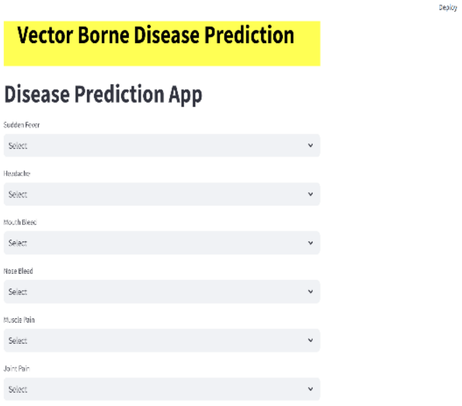
Logistic Regression doesn't require a direct linear relationship between variables**.**

3. FUTURE ENHANCEMENT:

Controlling vector-borne diseases in India is challenging due to diverse climatic conditions that increase outbreak risks. This research proposes a predictive model that diagnoses such diseases early and classifies their types, achieving 97.5% accuracy. The model aids medical teams in making timely decisions, saving lives. It can be extended to other diseases like chikungunya, Zika, and kala-azar using advanced machine learning techniques.

4.RESULT:

INPUT IMAGE:

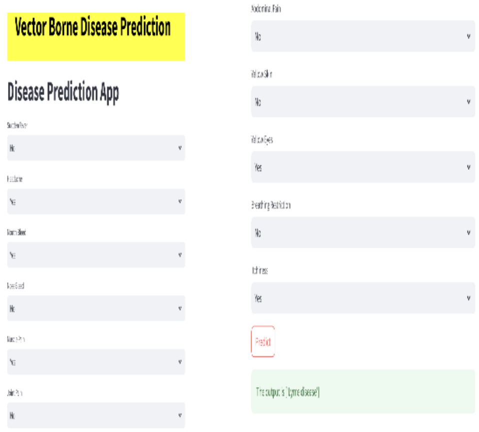


Vector Borne Disease Prediction

Disease Prediction App

Sudden Fever: Select
Headache: Select
Mouth Bleed: Select
Nose Bleed: Select
Muscle Pain: Select
Joint Pain: Select

OUTPUT:



Vector Borne Disease Prediction

Disease Prediction App

Sudden Fever: No
Headache: Yes
Mouth Bleed: Yes
Nose Bleed: No
Muscle Pain: Yes
Joint Pain: No
Abdominal Pain: No
Yellow Skin: No
Yellow Eyes: Yes
Breathing Restriction: No
Itchiness: Yes

The output is ['Lyme disease']

5.CONCLUSION:

Methods like Logistic Regression and Random Forest are crucial for forecasting the spread of diseases such as malaria, dengue, and chikungunya. Logistic Regression efficiently manages intricate, non-linear connections between variables, while Random Forest excels in pinpointing key risk factors driving disease transmission. These approaches significantly improve prediction accuracy and dependability. With these tools, health workers can respond faster, plan better, and use resources more effectively, saving more lives and reducing the impact of these diseases. This is especially important in countries like India, where these diseases affect a large number of people. Over time, improving these techniques and adding real-time data can make them even more helpful in controlling and preventing outbreaks.

6.REFERENCES:

[1] Dr. K. Usha Rani on "Analysis of heart diseases dataset using neural network approach" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011

[2] Shraddha Shivhare and Rajesh Shrivastava on "Automatic Bone Marrow White Blood Cell classification using Morphological Granulometric feature of Nucleus" Oriental Journal Of Computer Science & Technology Vol. 5, No. (1): Pgs. 75-85, June 2012

[3] Md. Osman Goni Nayeem, Maung Ning Wan and Md. Kamrul Hasan on "Prediction of disease level using multilayer perceptron of Artificial Neural Network for patient monitoring" International Journal of Soft Computing and Engineering (IJSCE) Volume-5 Issue-4: Pgs17-23, September 2015

[4] Sivaranjani. R and Dr .N .Yuvaraj on "Artificial Intelligence model for earlier prediction of cardiac functionalities using multilayer perceptron" International Conference on Physics and Photonics Processes in Nano Sciences

[5]Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou on "Lung pattern classification for interstitial lung diseases using a deep Convolutional Neural Network" IEEE Transactions On Medical Imaging, VOL. 35, NO. 5, MAY 2016

[6] Rahib H. Abiyev and Mohammad Khaleel Sallam Ma'aitah "Deep Convolutional Neural Networks for chest diseases detection" Journal of Healthcare Engineering Volume 2018

[7] Agus Qomaruddin Munir and Edi Winarko on Classification model disease risk areas endemicity Vector Borne Disease Fever outbreakbased prediction of Patients, death, IR and CFR using forecasting techniques" International Journal of Computer Applications, March 2015

[8] Juan M. Scavuzzoa , Francisco Truccoa , Manuel Espinosac , Carolina B. Taurob , Marcelo Abrilc , Carlos M. Scavuzzob ,and Alejandro C. Freryd Modeling on "Vector Borne Disease vector population using remotely sensed data and Machine Learning" in press.

[9] Mahalakshmi and G. Suseendran on Prediction of Zika Virus by multilayer perceptron neural network (MLPNN) using Cloud B" International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-2S11: Pgs. 249-254, September 2019

[10] V.Janani , N. Maadhuryaa , D. Pavithra and S. Ramya Sree on "Vector Borne Disease Prediction using multilayer perceptron - a Machine Learning approach" International Journal of Research in Engineering, Science and Management Volume-3, Issue-3, March 2020

[11] Oladimeji Mudele ,Fábio M. Bayer, Lucas Zanandrez, Alvaro E. Eiras and Paolo Gamba on "Modeling the temporal population distribution of Ae. Aegypti mosquito using big Earth observation data" in press.

[12] Vijeta Sharma ,Ajai Kumar ,Lakshmi Panat ,Dr. Ganesh Karajkhede and Anuradha lele on "Malaria outbreak prediction model using Machine Learning" International Journal of Advance

Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 12, December 2015