The Journal of Supercomputing (2024) 80:15725–15762
https://doi.org/10.1007/s11227-024-06094-w

# ROVM integrated advanced machine learning-based malaria prediction strategy in Tripura

**Apurba Debnath[1] · Anirban Tarafdar[1] · A. Poojitha Reddy[1] · Paritosh Bhattacharya[1]**

Accepted: 19 March 2024 / Published online: 6 April 2024


## Abstract

Malaria is a deadly disease that can take a person's life if not predicted or cured correctly. Numerous factors like temperature, humidity, precipitation, etc., impact India's increasing cases of malaria diseases. This research presents an advanced machine learning regression technique recently developed to anticipate the prevalence of malaria in Tripura using a real-world data. The proposed structure uses nine different regression methods, such as multilayer perceptron (MLP), random forest, support vector regressor, gradient boosting regressor, Bayesian ridge, kernel ridge, extreme gradient boost regressor (XGB), light gradient boosting machine (LGBM regressor and linear regression, to predict malaria using the most affecting factors of malaria diseases, namely temperature, humidity, precipitation, month, and years as input. Furthermore, to opt out the best suited technology for malaria cases prediction, the range of value method (ROVM)–multi-criteria decision methods (MCDM) technique has been applied, considering various statistical measurements as criteria. Ultimately, a comparison of various MCDM techniques reveals that MLP, XGB, and RF emerge as the top three choices. MLP regression, with root-mean-square error (RMSE) value of 0.03357, yields the lowest RMSE value, and the coefficient of determination ($R^2$) is 0.97616, yielding the maximum among other regressions. To effectively battle the illness in Tripura, it could be useful for continuing intervention tactics by governmental, profit and nonprofit organizations.

**Keywords** Prediction system · Regression model · Machine learning · Malaria case

✉ Anirban Tarafdar
anirban.ganit@gmail.com

[1] Department of Mathematics, NIT Agartala, Agartala, India

## 1 Introduction

The parasite disease malaria, which is spread by mosquitoes, can infect both humans and animals [1]. Fever, tiredness, and headaches are typical malarial manifestations. In the worst-case situation, it can result in death, unconsciousness, jaundice, or convulsions. Usually, warning signs appear within 10 to 15 days after receiving a bite by an infected mosquito. Many months after obtaining inadequate medical attention, the condition may return in some people. Pandemics of malaria are thought to have flourished long before civilization. The most prevalent illness, it has taken a great number of lives, and it has even been connected to substantial military defeats and the collapse of some nations. Despite declining malaria morbidity and mortality in some regions, it persists elsewhere due to deteriorating health services, drug and insecticide resistance, poor water management, and socioeconomic and environmental factors. Effective prevention requires simple methods for accurate forecasting, early warnings, and rapid case detection in both low- and high-transmission areas.

There is a close relationship between the dynamics of malaria transmission and environmental variables, including climatic and meteorological characteristics. Relative humidity, temperature, and precipitation variations all have a significant impact on Anopheles mosquitoes, which are the vectors that spread malaria parasites. Numerous elements of the mosquito's life cycle, such as reproduction, development, and survival rates, are influenced by these environmental factors. Additionally, they affect how the malaria parasite grows and spreads within the mosquito. The present study focuses heavily on understanding the link between precipitation, temperature, and humidity given the tremendous effect of climate on malaria transmission, particularly in tropical countries [2, 3] where malaria is prevalent. Our goal is to clarify how these climate factors influence the temporal and geographical patterns of malaria incidence by examining them. To facilitate a comprehensive understanding of the malaria transmission process, Fig. 1 illustrates the developmental stages of the malaria parasite within the mosquito vector.

Since malaria kills over 400,000 individuals annually throughout the world, scientists have been diligently studying its risk factors and core causes for years [4]. Since it was discovered that one of the primary risk factors for malaria is human genetics, numerous genome-wide association studies (GWAS) have been conducted. India has a billion people who could contract malaria. The propagation of diseases is challenging because of the diversified ecosystem, availability of transmission vectors, and biological features. In the northeastern state of India, where malaria is prevalent and continuously produces 10% of cases (the majority of those caused by Plasmodium falciparum) and 20% of confirmed fatalities each year, it is projected that 4% of the country's population resides. Figure 2 shows an illustration of the rise in malaria deaths in several nations during the past few years (2015–2021).

Tripura nestled in India's northeastern region shares an extensive border with Bangladesh, rendering it strategically positioned. Its diverse, landlocked terrain

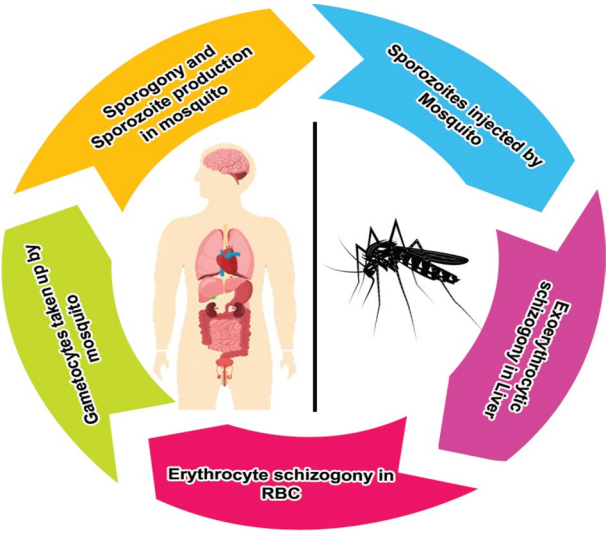


**Fig. 1** Represent schematically the development process of malaria parasite in human body

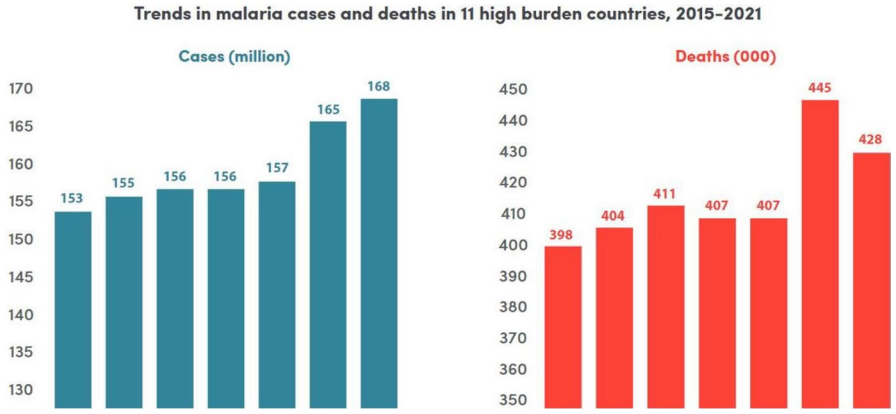


**Fig. 2** Trends in malaria cases and death from 2015 to 2021 [7]

hosts numerous ethnic communities with rich cultural practices. Given malaria's susceptibility to weather, meteorological data like precipitation and temperature are pivotal for tracking outbreaks. The proximity to Bangladesh, where Plasmodium falciparum drug resistance is prevalent, underscores Tripura's significance despite its small size. This poses a significant challenge for its 3.2 million residents, visitors, and border patrol. Consequently, early malaria detection and surveillance are imperative to safeguarding these populations amidst evolving medication resistance and border vulnerabilities.

Malaria cases in northeast Tripura are intricately linked to temperature, precipitation, and humidity due to the region's unique climatic conditions. As a vector-borne

disease, malaria transmission is greatly influenced by environmental factors. Elevated temperatures create favorable conditions for the breeding and survival of mosquitoes, the primary carriers of the malaria parasite [5]. Similarly, high levels of precipitation provide breeding sites for mosquitoes in stagnant water, further facilitating transmission. Additionally, humidity plays a crucial role in the survival and activity of mosquitoes, affecting their ability to seek hosts and transmit the disease. In northeast Tripura, where malaria is endemic, the combination of these climatic factors creates a conducive environment for the proliferation of malaria-transmitting mosquitoes and the spread of the disease [6]. Monitoring temperature, precipitation, and humidity patterns becomes essential for predicting malaria outbreaks, enabling timely interventions such as vector control measures and healthcare provision to mitigate the disease burden in the region.

This study presents a novel machine learning method to forecast malaria trends, aiming to improve understanding of its dynamics and enhance healthcare management. The historical data are collected from the governmental hospital, and climate data have been gathered from official site. By analyzing historical data with advanced computing, it seeks to inform proactive prevention and control strategies, offering insights for targeted public health initiatives. Integrating machine learning for disease prediction enhances prognosis accuracy and equips healthcare professionals with tools for societal benefit through preventive actions. The research aims to bridge knowledge gaps and address model limitations by developing a machine-learning-based malaria prediction model for Tripura, focusing on heavy equipment residual value prediction. It critically evaluates existing models' shortcomings through comprehensive analysis of factors affecting equipment's remaining value. Employing diverse machine learning techniques, the study crafts a reliable, innovative model, enhanced by meticulously organized datasets for improved accuracy.

Furthermore, it utilizes advanced multi-criteria decision-making (MCDM) methods, specifically the range of value method (ROVM), to identify the most effective regression technique for accurate predictions. The range of value method (ROVM) approach is used in the present study to forecast illnesses using a dataset on malaria. Because ROVM is dependent on decision-makers' ordinal evaluations of criterion relevance, it is especially useful in situations where determining numerical weights might be difficult [8]. Regarding MCDM, ROVM streamlines ranking options based on multiple criteria, aiding in the identification of optimal malaria prediction outcomes. Decision-makers can compare choices within specified value ranges for each criterion, facilitating informed decision-making. ROVM offers comprehensive evaluation of multiple elements simultaneously, enhancing transparency and facilitating trade-off analysis. In manufacturing [8], ROVM simplifies cutting fluid selection, proving effective across various case studies. Its simple computation and adaptability, capable of handling numerous criteria without complexity, are notable advantages. Additionally, when combined with fuzzy scales, ROVM accommodates qualitative criteria, enabling subjective assessments in decision-making.

In the proposed study, upon obtaining machine learning results, it became evident that the statistical errors for various regression techniques exhibited striking similarity. This made it challenging to discern the most optimal ML regression technique for malaria case prediction. To address this, the study employed range

of value method (ROVM) multi-criteria decision-making (MCDM) techniques to determine the best approach.

### 1.1 Research objectives

The paper concentrates on early prognosis of malaria to facilitate timely treatment. After thorough research, we developed nine regression-based predictive models, aiming to improve prediction accuracy and robustness. As malaria prevalence increases, creating accurate predictive models is crucial for timely intervention and saving lives by preventing the spread of the disease. Six basic steps make up the present modeling process. These are discussing below:

- Data from 60 typical datasets for each type of targeted heavy equipment were collected from Dharmanagar Hospital in step 1. Meteorological conditions were also considered to predict diseases. The data underwent identification, training with regression algorithms, and the results were imported into MS Excel spreadsheets.
- Pre-processing of data was carried out before each phase in the framework's creation process to increase the predictive potential of the model being created.
- The third step in the process is feature selection and identifies the characteristics that have the greatest influence on heavy construction equipment's residual value.
- Modeling and comparison using machine learning in step 4. The precision of each algorithm is gauged by comparing observed residual values to estimated ones. Four evaluation metrics—MAE, RMSE, MSE, and MAPE—were utilized to identify the most reliable model.
- Utilizing the MCDM method, specifically the ROVM, enables the identification of the optimal regression technique for disease prediction.
- Through the comparison and evaluation of modeling results, output depiction and conversation give users a hierarchical perspective of the reliable approach to the residual value estimation for malaria disorders.

### 1.2 Novelty

Extensive research has been dedicated to the study of malaria. Nevertheless, the literature review in Sect. 2 indicates that the accuracy or predictive performance of existing methodologies is considerably low and falls short in effectively addressing real-world situations. Additionally, practically little is being done in Tripura. Numerous considerations must be considered while selecting the best machine learning algorithm for disease prognosis. MCDM techniques provide systematic ranking based on criteria like accuracy and interpretability, enhancing decision-making and illness forecasting accuracy in medicine:

- The first-ever case study conducted in Tripura's north district centers on malaria, seeking to unveil insights for precise interventions and enhance public health outcomes.
- The present study involves an analysis based on real datasets collected from Dharmanagar District Hospital. Also, several meteorological data have been collected for Dharmanagar district Tripura.
- Sophisticated machine learning methodologies have been implemented to thoroughly analyze a dataset obtained from real-world sources.
- Six statistical metrics were evaluated in order to examine errors thoroughly and provide an extensive overview of the quality and reliability of the dataset.
- The study employed the latest MCDM technique, ROVM, to identify the most effective forecasting method, offering advanced comparison across various parameters. It aimed to determine the optimal machine learning method for early disease prediction through ROVM.
- The study utilized statistical metrics such as RMSE, MSE, R, R^2, MAPE, and MSRE to construct a decision matrix within the MCDM technique. This aided in determining the most suitable among the nine models considered for prediction.
- The study conducts a comparison between the ranking generated by the ROVM and various other MCDM techniques. Through correlation analysis, it validates the identification of the most effective machine learning approach for predicting diseases at an early stage..

## 2 Literature review

The associated research that was done in order to create the suggested model for malaria disease prediction is described in this part. The comment that is made below is based on an analysis of the literature that contributes to the efficient and effective development of the suggested system.

Fast diagnosis is essential for managing malaria. The use of blood smear pictures to train machine learning algorithms to identify malaria has been attempted in several researches, but this approach has serious flaws [9]. It uses the ML approach for diagnosing the malaria diseases using the patient's information. The study [10] suggests a brand-new, scalable approach to forecast malaria cases in particular regions. Long short-term memory (LSTM) classifier, medical information, and satellite data were utilized to forecast malaria distributions in the Indian state of Telangana. The paper [10] uses the data on malaria and make prediction by using big data. The study creates [11] a weather-based malaria prediction model and applies it to seasonal forecasts for climate employing weekly time-lapse data from Vhembe, Limpopo, South Africa, during 1998 to 2015 that includes precipitation, temperature, and malaria cases. The paper [12] compares the different supervised a disease detection algorithm using machine learning. They use five machine learning methods to compare the predictive values of the diseases. According to the study [13], Machine learning methods were used to predict heart problems with accuracy. [14] examined the statistical evaluation and creation of models for the use of ML for predicting the severity of the COVID-19 disease from clinical blood test data. The study uses

machine learning techniques in COVID-19 data. Many immunological mediators have significant roles in the etiology of illnesses affecting the eye. To create models with the greatest amount of predictive value, various predictors can be dynamically selected and ranked using machine learning [15]. A total of five ML approaches were used in [15]. Since endemic malaria is present in many sub-Saharan African countries, including the countries of Burkina Faso, Mali, Republic, Niger, Nigeria, Democratic Republic of Congo (DRC), and Cameroon, machine learning has the potential to be used to independently choose and weigh different predictors [16].

Climate variability is the term used to describe variations in environmental factors such as precipitation, humidity, air temperature, and atmospheric pressure [17]. The study [6] has created and evaluated a statistical model of thelinks between cases of malaria and climatic factors. The protozoan parasites of the genus Plasmodium are what cause malaria. Plasmodium Malariae, Plasmodium falciparum, Plasmodium vivax, and Plasmodium ovuleare the parasites that cause it in people. P. falciparum is the most prevalent of them, causing 80% of all cases of malaria infections and 90% of fatalities in Africa and South East Asia [18]. The study [19] uses the techniques of machine learning to the chronic illness of a Patient. By utilizing machine learning algorithms like convolution neural network, or CNN, for automatically obtaining features and forecasting disease and K-nearest neighbors (KNN) for determining distance with the goal to identify the precise correspondence in the information set and calculate the final disease estimation results, the suggested approach provides a broad outlook for disease based on patient symptoms. The study [20] employed a machine learning approach with variation in the climate for forecasting malaria incidents. This study suggests a machine learning-based methodology for categorizing malaria prevalence using climate variability during a 28 year period in six sub-Saharan African nations. The study [21] focuses on the prediction of air quality index in Kolkata city using an advanced learned interval type-3 fuzzy logic system.

The study [22] uses advanced mathematical modeling and artificial neural network techniques to reduce inventory costs amidst COVID-19 challenges. The study [23] aims to enhance early breast cancer detection using artificial intelligence (AI) and machine learning) ML algorithms. It evaluates eight regression models to select the most accurate predictor. The study proposes an AI-based system for early diabetes prediction, utilizing Ridge-Adaline SGD classifier [24]. The study [25] offers a superior deep-learning-based secure block chain (ODLSB) enabled Internet of Things (IoT) and healthcare diagnosis model. The research gap table is shown in Table 1.

### 2.1 Research gap

After further investigation, it becomes clear that the predictive methods for predicting cases of malaria have not been fully employed. What is also remarkable is how little study has been done in Tripura on malaria prediction. The lack of research exposes a serious gap in our knowledge of and ability to address the dynamics of malaria in the area. Effective disease management and preventive measures depend

**Table 1** Research gap of literature review

| Sl. num | Objective | MCDM application | Tripura | Malaria | ML` |
|---|---|---|---|---|---|
| 1 | Machine learning model for predicting malaria using clinical information [1] | No | No | Yes | Yes |
| 2 | LSTM-based prediction of malaria abundances using big data [2] | No | No | Yes | No |
| 3 | Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model [3] | No | No | Yes | Yes |
| 4 | Comparing different supervised machine learning algorithms for disease prediction [4] | No | No | No | Yes |
| 5 | Effective heart disease prediction using hybrid machine learning techniques [9] | No | No | No | Yes |
| 6 | Machine learning techniques for pavement condition evaluation [10] | No | No | No | No |
| 7 | Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development [11] | No | No | No | Yes |
| 8 | Machine learning approach for intraocular disease prediction based on aqueous humour immune mediator profiles [12] | No | No | No | Yes |
| 9 | Identification and prediction of chronic diseases using machine learning approach [13] | No | No | No | Yes |
| 10 | Prediction of malaria incidence using climate variability and machine learning [14] | No | No | Yes | Yes |
| 11 | Our Model | Yes | Yes | Yes | Yes |

on more research and the creation of prediction models specific to Tripura's unique features. Comparison of different model is important to find which model is best for the prediction. This paper has included a total of nine regression methods, which is very huge that no other studies have study have taken that much numbers of techniques.

Numerous environmental parameters, such as temperature, humidity, and precipitation, have been thoroughly analyzed in the current study. To better understand the roles, they play in influencing the dynamics of malaria transmission, and these factors have been in-depth analysis. In order to contribute to more efficient methods for disease prevention and control, the study considers various climatic factors in order to offer a thorough knowledge of the complex link between environmental circumstances and the incidence of malaria.

It has been observed that in some studies malaria cases were consider, but regression techniques have been not used. In some paper, malaria and machine learning approach were used but climate factor has not been considered. In some studies, malaria, machine learning, and climate factor considered, but the ranking procedure was not considered. But the current study includes all the required things, i.e., malaria, machine learning, and MCDM techniques so that malaria can be predicted more accurately and early cure can be taken based on this prediction. The current study has included range of value method (ROVM) MCDM techniques to do the ranking and to find the optimal or the best method for predicting the malaria in Tripura.

## 3 Proposed framework

The methodical approach used in this study is outlined in the proposed framework which is explained in this section. First, information gathering from the Hospital was initiated, including obtaining unprocessed datasets. After then, these datasets were rigorously arranged to give priority to relevant data. This organized configuration allowed for the identification of critical inputs for analysis, which included variables like temperature, humidity, precipitation, month, and year. Following that, the framework gave a machine learning approach that used nine different regression techniques, as seen by the flow chart in Fig. 3. All these methods were subjected to extensive statistical analyses to see how well they modeled the dataset. After doing this thorough assessment, the best model was identified by applying multi-criteria decision-making (MCDM) methods. Through this selection procedure, the researchers were able to make sure that the selected model met the predetermined criteria and properly reflected the dataset.

The study's structural arrangement was described as follows: Section 4 included an overview of the sources and techniques used in the data collecting process. The detailed elements of data organization, input identification, output specifications, and the methods used for prediction were covered in Sect. 5. Section 6 covered further detail, examining a wide range of machine learning methods used in the framework. This part offered insights into each algorithm's theoretical foundations and suitability for use in the study's environment along with Process simulation, where

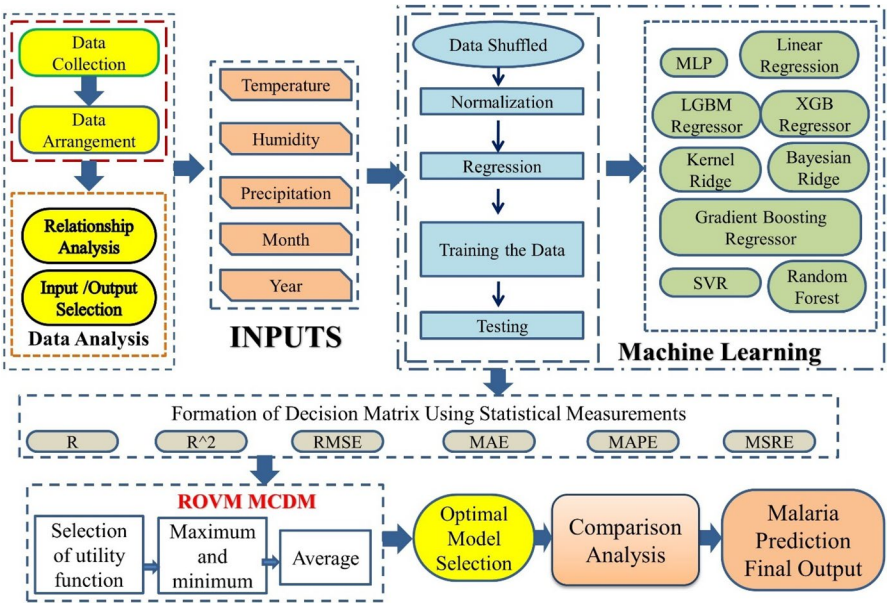


**Fig. 3** Flowchart of the proposed model

real-world scenarios were simulated using the suggested framework. This made it possible to verify the framework's usefulness in real-world settings. In Sect. 7, results have been discussed. The study ended with a thorough analysis of the results, offering insightful information on the implications and possible future paths of the investigation.

## 4 Data collection

A thorough explanation of the given dataset construction, building of model, and disease estimations have been provided in this section. Data collecting comes first in the process. Data from both organized and unstructured sources are gathered by our suggested system. Datasets are divided into cleaning and test datasets when preprocessing is applied to the collected data. Monthly malaria cases along with the number of tests were collected for this project.

### 4.1 Data collection process

Data on monthly malaria case counts were gathered from the Chief Medical Officer, District Hospital, North Tripura district of Tripura (2018–2022) in order to understand the malaria epidemiology. The study gathered information on the total the number of malaria cases throughout the five years before (2018–2022) in the North Tripura district of Tripura. There are 60 distinct datasets in the study of

present paper. These data are in the form of month wise number of malaria cases including the number of tests that were used to check whether a patient have the diseases or not. One of the premier medical facilities in the district of North Tripura is Dharmanagar Hospital. This prompted more people to seek therapy here. To ensure that the data obtained are authentic and trustworthy for upcoming statistical analysis, the data for this research were picked from Dharmanagar Hospital as the venue for data gathering. These statistics showed the percentage of patients with fever who had malaria cases among all patients. The number of people who were tested and determined to have malaria was included in the hospital data that was compiled at the level of local administrative units, health centers, and districts. Figure 4 offers a detailed depiction of the location and interior of the Dharmanagar district hospital. This visualization aims to underscore the importance of this hospital within the context of Tripura, providing readers with a comprehensive understanding of its role and relevance.

Monthly reported cases of malaria caused by P. falciparum and P. vivax make up the epidemiological statistics. By using immune-chromatography-based rapid diagnostic tests (RDTs) and microscopic analysis of blood smears during the monitoring (active/passive), the malaria cases were identified.

Different temperature extremes were recorded during the five years, from January 2018 to December 2022, providing important information on the climatic differences that occurred during this time. The maximum recorded temperature during the time under scrutiny was 41 °C in April of 2018. This indicates a particularly warm season. On the other hand, February 2021 experienced a sharp drop in temperatures, with the mercury falling as low as 8 °C. These observed values provide important insights into the climate dynamics that prevailed throughout the designated period, illuminating seasonal variations and possible anomalies.

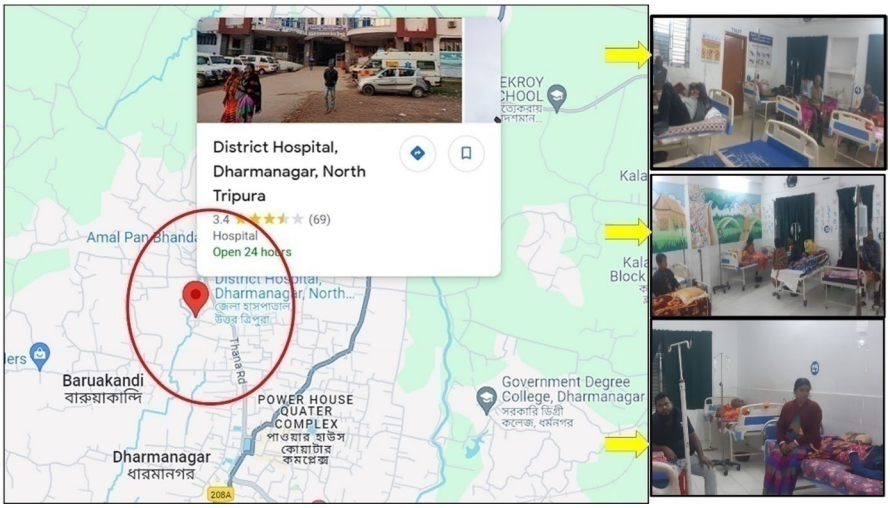


**Fig. 4** District hospital, Dharmanagar, North Tripura, Tripura

The previous five years of data of North Tripura's monthly maximum and minimum temperatures (January 2018–December 2022) and the precipitation were gathered from [26]. Precipitation is the process through which water vapors condenses in the surrounding air and falls to the ground as a result of gravity. The five most recent years (January 2018–December 2022) of North Tripura's monthly humidity data were gathered from [27].

## 5 Data perception

Although monthly malaria case data have been collected for this study, more knowledge about the illness is required to improve prediction accuracy. Climate-related elements, including temperature, precipitation, and humidity, have a significant impact on malaria prevalence, which frequently increases in tandem with these environmental changes. Thus, in order to get more accurate forecasts, it is essential to incorporate these climatic factors into the predictive model. The correlation between malaria incidence and climate fluctuations, namely in temperature, humidity, and precipitation, has been established. Consequently, it is apparent that the prevalence of the illness varies considerably across different regions. Therefore, taking into consideration these environmental dynamics is crucial to capture the intricate patterns of malaria transmission and guaranteeing the accuracy of prediction models in various geographical locations.

The monthly data only temperature for the last five years (2018–2022) were in the Celsius scales. So, it can be said that the number of cases increases with the increase in the temperature. Tripura's humidity is often high all year round. The relative humidity varies from 50 to 74% in the summer, but it exceeds 85% during the wet season. By applying a machine learning approach, we can obtain the predicted value of malaria diseases in the output.

### 5.1 Inputs

It is critical to recognize and include pertinent input factors while creating a machine learning algorithm for malaria illness prediction. It is crucial to collect data that represents the factors influencing malaria prevalence because the focus of this study is malaria prediction. The prevalence of malaria varies regionally as a result of variables such seasonal variations in temperature, humidity, and precipitation patterns. The variables are therefore chosen to serve as the machine learning algorithm's inputs. Variations in temperature affect the growth and survival of mosquitoes, which are the primary vectors of the malaria parasite, and hence have a major impact on the dynamics of malaria transmission. In a comparable manner, excessive humidity increases mosquito breeding grounds and therefore the risk of malaria transmission. The distribution of malaria is further influenced by the ways in which precipitation patterns impact mosquito breeding grounds.

- *Temperature* Temperature has a vital effect in the increases in the malaria cases. It has been observed that the maximum temperature of the state is in between 37 and 41 °C. The minimum temperature of the state lies in between 10 and 15 °C. Additionally, it has been established that, over the past five years, the months of July to November have had the highest occurrence of malaria across the majority of North Tripura. The following graph in Fig. 5 shows the past five years (2018–2022) month-wise collection of temperature of Tripura. Here in horizontal axis of the graph, number of months starting from January 2018 at the point 1 and ending up to December 2022 at the point 60.
- *Humidity* Tripura consistently has 60–80% humidity, which is perfect for Anopheles mosquito reproduction. The ideal conditions for malaria transmission are at least 60% humidity and 18–32 °C, which these insects like. The monthly humidity trends for Tripura from 2018 to 2022 are shown in Fig. 6, which makes it possible to evaluate the danger of malaria and mosquito activity. By keeping an eye on these patterns, one may successfully prevent the spread of malaria by anticipating illness outbreaks and directing the adoption of focused control measures. In order to prevent malaria and protect the public health of the area, it is essential to comprehend environmental elements, especially humidity.

- *Precipitation* The pattern of precipitation, as seen in Fig. 7, is a significant factor that impacts the incidence of malaria cases in a particular region. Changes in the amount of precipitation have a direct effect on the mosquito breeding grounds that transmit malaria, which in turn affects the dynamics of the disease's transmission. The graph shows the monthly patterns in precipitation that Tripura has

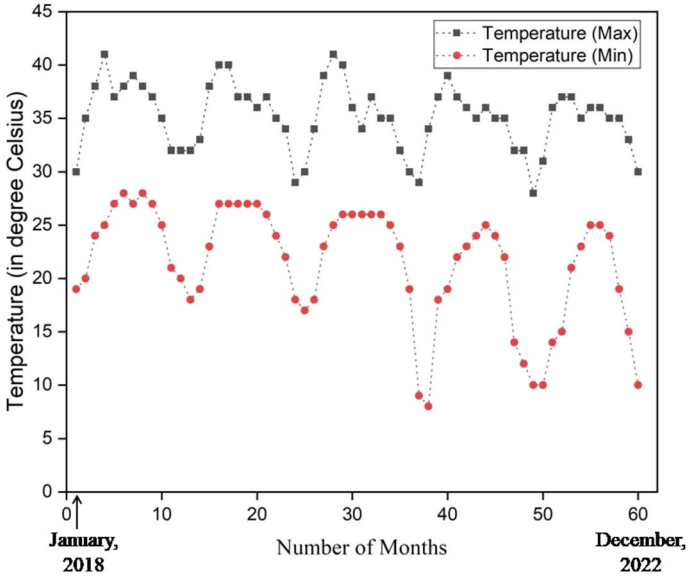


**Fig. 5** Month-wise variation of temperature for the past 5 years (January 2018–December 2022)

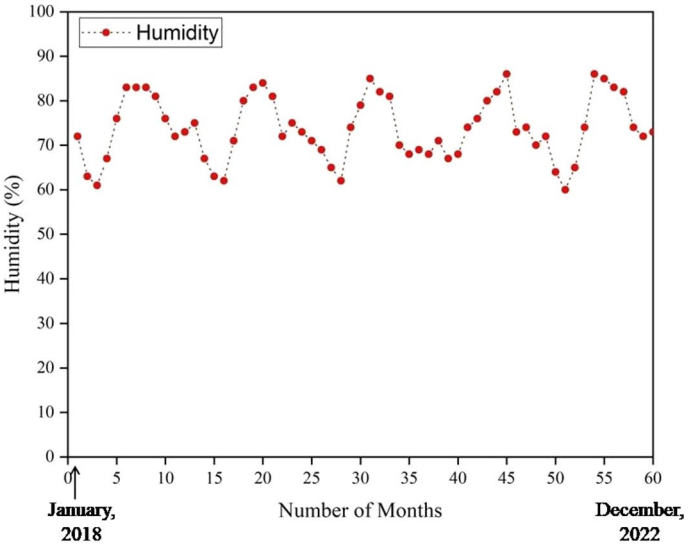


**Fig. 6** Month-wise variation of humidity for the past 5 years (January 2018–December 2022)

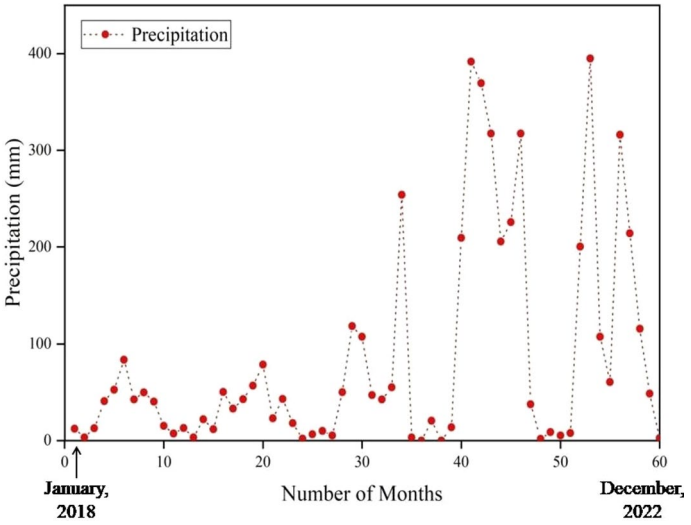


**Fig. 7** Month-wise variation of precipitation for the past 5 years (January 2018–December 2022)

seen from January 2018 to December 2022, a period of five years. The seasonal variations in rainfall that can greatly affect mosquito breeding patterns, and, in turn, malaria incidence rates are revealed by these precipitation data.

- *Months* All months from January 2018 to December 2022 are considered as one of the inputs of the model. As the variable changes significantly with respect to different months; therefore, months are one of the crucial inputs. Initially in the

collected data, months are in linguistic terms but in the calculation process linguistic terms are converted to the number format 1 to 12.

- *Year* The analysis's input data are available for five years, from January 2018 to December 2022. This period is broken up into distinct chunks, each of 12 months. The input data are taken to correspond to the year 2018 for the first 12 month period, which runs from January to December 2018. The next twelve months, from January to December 2019, are then classified as input data for the year 2019, and so on for the years that follow. This way of arranging the input data helps to keep the analysis consistent and makes it easier to look for trends and patterns throughout the course of the five years.

### 5.2 Output

The major output of the present paper is to focus on the prediction of malaria diseases of Tripura state.

- *Malaria cases* For better results, the paper has also included the past five years (2018–2022) malaria cases of North Tripura district. Month-wise number of malaria cases was collected. These data were divided into the number of affected individuals by the Plasmodium falciparum and Plasmodium vivax. The data also show the number of death and recovered people at that period.

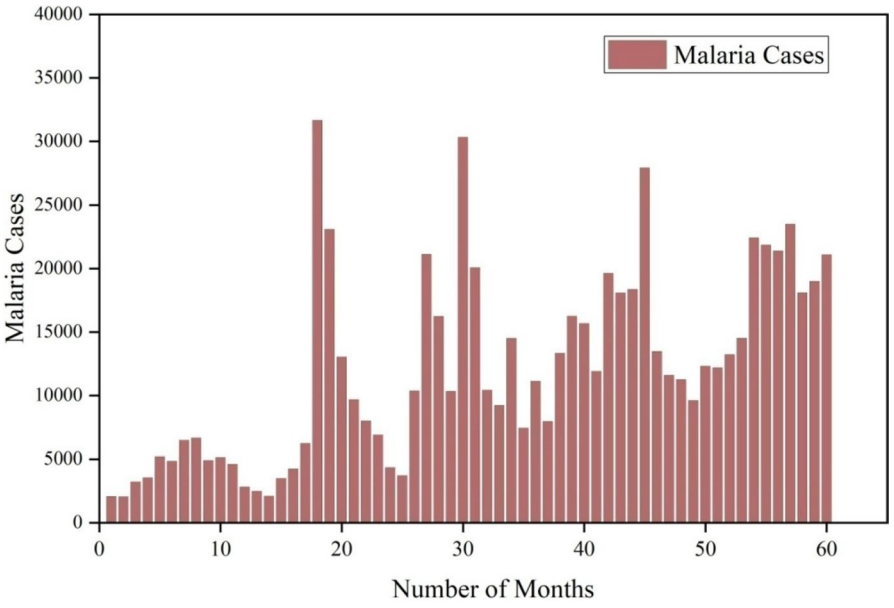


**Fig. 8** Month-wise collection of number of cases of malaria for the past five years (January 2018–December 2022) of North Tripura district

Figure 8 shows the month-wise collection of number of cases of malaria for the past five years (2018–2022) of North Tripura district.

### 5.3 Correlation between input and output

The two primary advantages of correlation analysis are that it enables quick hypothesis testing and assists businesses in deciding which variables they wish to investigate further. Here, for basic understanding among the inputs and outputs components, different 3D graphs have been used. The correlation coefficient aids in determining how closely two variables are related in a single figure.

Figure 9 illustrates the relationship between temperatures, the duration of months, and precipitation. This visual representation allows for the observation of any patterns or trends between these variables over time. Meanwhile, Fig. 10 depicts the correlation between the input variables, namely humidity and the duration of months, and the output variable, which is the number of malaria cases. This graph enables the examination of how changes in humidity and the duration of months may impact the incidence of malaria cases, providing insights into potential factors influencing disease prevalence.

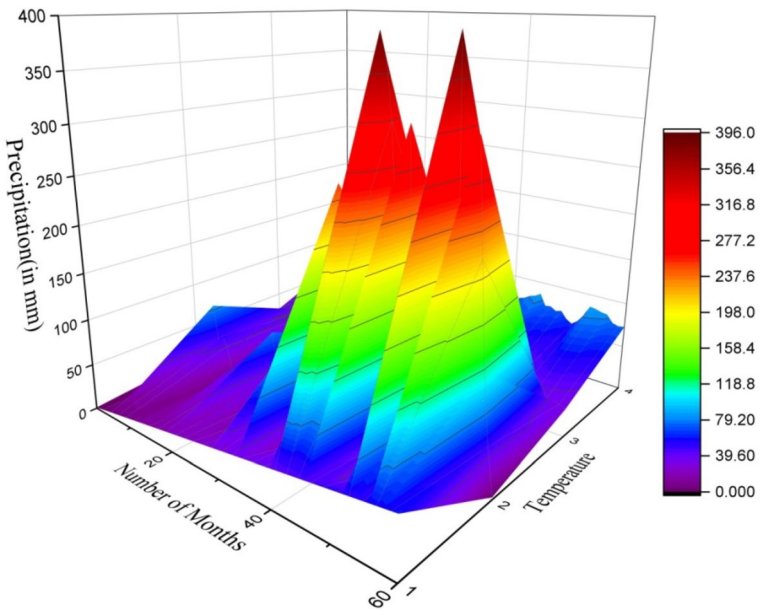


**Fig. 9** Correlation between temperature, number of months, and precipitation

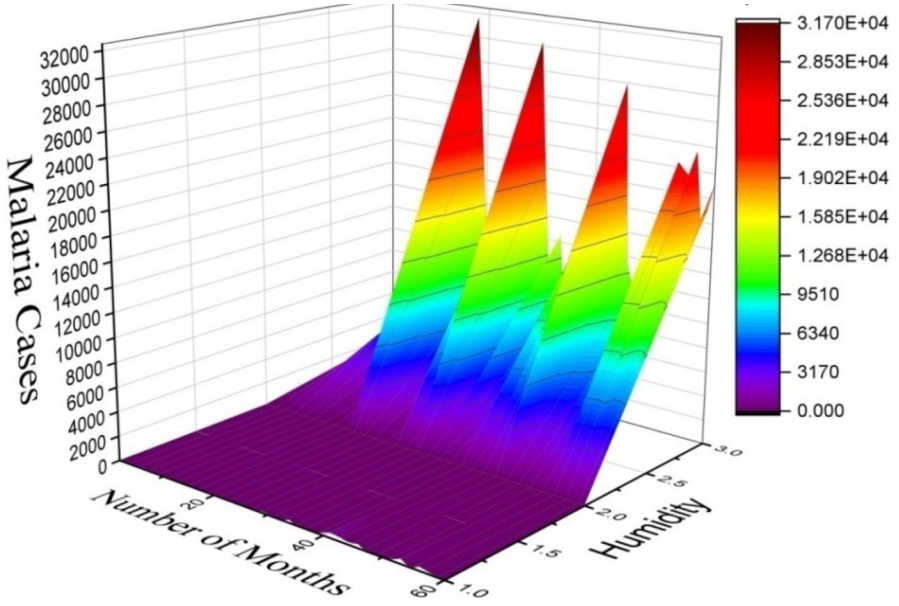


**Fig. 10** Correction between humidity, number of months, and number of malaria cases

## 6 Prediction methodology

The methodology unfolds in three sequential tasks: firstly, the construction of machine learning techniques; secondly, simulating the malaria model using the collected dataset and formulating the decision matrix based on the predicted output; and lastly, employing the ROVM MCDM technique to sustain the prioritization of the machine learning models. This systematic approach begins with technique development, progresses to model simulation and decision matrix formulation, and concludes with the application of the ROVM MCDM technique for model prioritization, ensuring a comprehensive and methodical analysis throughout the research process.

### 6.1 Machine learning algorithms

This study discusses the fundamental theory behind various regression algorithms used, which effectively match different classification methods. Each technique can capture complex interactions between input features, providing understandable results. Prediction, a powerful supervised ML method, utilizes independent and historical data to uncover patterns and correlations with output variables, forecasting new data accurately [28]. By learning from past instances, it identifies crucial data features indicating desired outcomes, encompassing behavioral patterns, environmental factors, or demographic information, depending on the domain.

To achieve the highest accuracy, it is imperative to utilize new prediction algorithms renowned for their adeptness in managing interconnected facets and handling intricate variable relationships. Nine regression techniques, including multilayer perceptron (MLP), random forest (RF), support vector regressor (SVR), gradient boosting regressor (GBR), Bayesian ridge (BR), kernel ridge (KR), extreme gradient boost (XGB) regressor, light gradient boosting machine (LGBM) regressor, and linear regression (LR), were employed in the current study. These techniques are explained below.

#### 6.1.1 Multilayer perceptron (MLP)

Regression can be carried out using the neural network known as the MLP regressor. Regression in machine learning can be conceptualized as a mapping from one space to another, each space having a different number of dimensions. The full form of MLP stands for "multilayer perceptron." This neural network is trained using supervised learning to predict output data points based on input data points by supplying input and output data as datasets [29]. The activation function of MLP Regressor is provided below:

$$y(v_i) = (1 + e^{-v_i})^{-1} \text{ and } y(v_i) = \tanh(v_i) \tag{1}$$

The first is a logistic function that has a similar form to the first but has a range from 0 to 1, while the second is a hyperbolic tangent that goes from $-1$ to 1. Here, $v_i$ is the weighted sum of the input connections, and $y_i$ is the ith node's (or neuron's) output.

#### 6.1.2 Random forest (RF)

A random forest is a kind of the statistical estimator that, in order to improve projected accuracy and minimize over fitting, takes the average of numerous classification decision trees that have been fitted to various dataset subsamples [30]. To achieve the optimum result, an ensemble classifier constructs multiple decision trees and combines them. Bootstrap aggregating or bagging is mostly used for tree learning. When given a set of data, $X = \{x_1, x_2, x_3, ..., x_n\}$ and the corresponding responses $\{Y = x_1, x_2, x_3, ..., x_n\}$ repeat the bagging process from $b = 1$ to B.

The scene sample $x\prime$ is made by arranging the predictions $\sum_{b=1}^{B} fb(x\prime)$ from every individual tree on $x\prime$:

$$j = \frac{1}{B}\sum_{b=1}^{B} fb(x\prime) \tag{2}$$

The uncertainty of prediction on this tree is made through its standard deviation:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(fb(x\prime) - f)^2}{B-1}} \tag{3}$$

### 6.1.3 Support vector regression (SVR)

The support vector regressor, or SVR for short, is an improvement on support vector machines, a machine learning (ML) technique for addressing regression issues. Combining statistical machine learning (ML) and structural risk minimization (SRMP), the SVM was developed [31].

Regression can be done using the support vector machine method and one of its core features i.e., maximum margin. The support vector regression (SVR) for classification adheres to the same principles as the SVM, with a few minor exceptions. First off, because outcome is a real number, there are an infinite number of alternative outcomes, making it very difficult to predict the information that is now available. Regression involves setting a tolerance margin (epsilon) that is generally in line with the problem's requirements that the SVM would have already specified. The algorithm is more complicated, so there is also a more complicated rationale that needs to be considered. The example below shows how SVR performs its principal function in Eq. (4).

$$\text{MinimiseSVR} = \frac{1}{2}\text{ww}^T + c\sum_{i=}^{N}\left(\alpha^+ + \alpha^-\right) \tag{4}$$

where $c$ and $\alpha$ were the SVR's parameters.

### 6.1.4 Gradient boosting regressor (GBR)

For classification and regression issues, a technique called gradient boosting is utilized in machine learning. It offers a model for forecasting is composed of weak prediction models that approximate decision trees [31, 32]. When a decision tree is used as the weak learner, the resulting approach, known as gradient-boosted trees, frequently beats random forest. Gradient-boosted trees models are constructed in the same stage-by-stage fashion as prior boosting procedures, but it generalizes those techniques by making any differentiable loss function optimization possible.

The field of learning to rank can benefit from gradient boosting. In high energy physics, gradient boosting is also used for data processing. Variants of gradient-boosting deep neural networks (DNNs) were successful in recreating the findings of non-machine learning methods of analysis on datasets used to detect the Higgs boson at the large Hadron collider (LHC).

The next equation gives the mathematical description of the GBR framework:

$$G_m(x) = \sum_{j=1}^{J} b_{jm}E(x \in R_{jm}) \tag{5}$$

here, $E\left(x \in R_{\text{jm}}\right) = \begin{cases} 1, if\, x \in R_{\text{jm}} \\ 0, otherwise \end{cases}$.where $b_{\text{jm}}$ is the overall number of instances in region $j$ and $R_{\text{jm}}$ is the total gradient in region $j$.

### 6.1.5 Bayesian ridge (BR)

Bayesian regression permits a natural mechanism to survive a lack of adequate data or data with an uneven distribution by employing statistical distributors rather than point estimates when producing linear regression. Rather than being computed as a single number, the output or answer '$y$' should be drawn from a probability distribution. First, we must recognize that Bayesian is just a strategy for developing and accessing statistical models. When there is insufficient or poorly dispersed data in a dataset, Bayesian regression can be highly effective. The outcome of a Bayesian analysis of variance is obtained from a distribution of probabilities, as opposed to conventional regression techniques, where the output is simply derived from a single value of each attribute. "$Y$" results from a normal distribution with normalized mean and variance. Instead of focusing on the model parameters directly, Bayesian linear regression seeks to discover the 'posterior' distribution for the model parameters [33, 34]

$$\text{Posterior} = \frac{\text{Likelihood} + \text{prior}}{\text{Normalization}} \tag{6}$$

When an event, such as $H$, occurs and another event, such as $E$, also occurs, the probability that $H$ will also occur is said to be posterior, or $P(H|E).P(H)$ tends for priority, which is the probability that event $H$ occurred before event $A$. Likelihood is a function in which a marginalized parameter variable is used.

### 6.1.6 Kernel ridge (KR)

Ridge regression and classification are combined with the kernel trick in kernel ridge regression (KRR). It then trains a linear function in the area provided by the suitable kernel and the data that is accessible. For irregular kernels, this is equivalent to a nonlinear function in the original space. Kernel ridge learned the support vector regression (SVR) model, and it takes the same structure. The kernel ridge is given by the function as below [35]

$$f := \left| \begin{array}{l} \text{argmin} \\ f \in H \end{array} \right. \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( f\left(x_i\right) - \left(y_i\right) \right)^2 + \lambda ||f||_{\mathcal{H}}^2 \right\} \tag{7}$$

where the regularization parameter is $\lambda > 0$. The estimator shown above is referred to as the kernel ridge regression estimate, whenever a replicating kernel Hilbert space is $H$. It is a natural extension of the ordinary ridge regression estimate to the nonparametric scenario (Hoerl and Kennard, 1970).

### 6.1.7 Extreme gradient boost (XGB) regressor

A recently created machine learning method called XGB regressor has been successfully applied in numerous sectors. Being a well-organized, transportable, and

adaptable approach, it will work for many different applications. Gradient boosting machine (GBM) is a method which integrates the cause-based decision tree (CBDT) with the gradient boosting machine (CBDT) into a powerful algorithm, enhancing the tree boosting approach to handle almost all data types rapidly and reliably. When regression and classification methods are used on the specified dataset, this algorithm's unique properties make it a useful tool for building forecasting models. Another use for XGB Regressor is for the processing of huge datasets with a substantial number of attributes and classifications [36].

#### 6.1.8 Light gradient boosting machine (LGBM) regression

Another advanced technology machine learning-based technique, LGBM (light gradient boosting machine regression), is used for data processing to do more accurate residual value modeling and prediction. The LGBM is used because of its exceptional competency, accuracy in data classification, and regression with a reasonably quick processing time. Exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS), two revolutionary data sampling and classification approaches, were combined to create this newly created technology. With such coupled characteristics, data scanning, sampling, clustering, and classification procedures are carried out swiftly and precisely compared to the analogue's techniques [37]. The loss function, often known as the pinball loss or quantile loss, distinguishes quantile regression from general regression. Here is a precise justification of the pinball loss using the following equation:

$$L\tau(y,z) = \begin{cases} (y-z)\tau, if & y \geq z \\ (z-y)(1-\tau), if & z > y \end{cases} \tag{8}$$

When $z$ is the targeted quintile and represents forecasted value, $y$ is the actual value.

#### 6.1.9 Linear regression (LR)

For predictive analysis, linear regression is the most widely used basic regression model. The first variable, a descriptive or explanatory variable, represents one of the two variables utilized in the model known as LR. A factor that is dependent or reliant variable is the second one [38].

The following is a representation of the standard set is the general single-equation linear regression model including the complementary subsets of simple regression and multiple regressions.

$$B = p + \sum_{i=1}^{m} q_i Z_i + u \tag{9}$$

where $B$ is the dependent variable, $Z_1, Z_2, ...Z_i...Z_m$ are m independent variables, $p$ and $q_i$ are the coefficient of regression, representing the model's parameters for a particular population, and u is a phrase for stochastic disruption that could be

explained by the influence of unspecified independent variables or by a completely random component in the relationship mentioned.

### 6.2 Process simulation and statistical measures

Processing simulation is the modeling-based portrayal of technological procedures and unit activities in software, including chemical, physical, biological, and other processes. The program's ability to determine process parameters depends on the physical and chemical characteristics of reactions, simple elements and mixtures, and computer models [39]. In this section, we will go over the methodology that has been used to analyze the malaria predictive model in North Tripura district of Tripura. The monthly prevalence of malaria over the previous 5 years (2018–2022) has been compiled. A total of 60 datasets were gathered from the Dharmanagar Hospital at the North Tripura district hospital. Now the data are shuffled to get the better experienced results. The data are then normalized using the normalization method, and the value of the data is transformed in between 0 and 1. In the machine learning, first the data has been shuffled, and then the data are normalizing to minimize the outlier.

#### 6.2.1 Simulation

The study utilizes the regression approach to predict the spread of malaria. The regression model was tested using 20% of the collected samples, while the remaining 80% were utilized to train the regression model. Several well-known ML models, including LGBM (light gradient boosting machine), gradient booster, XGB regressor, linear regression, SVR (support vector regression), MLP (multilayer perceptron), etc., were put to the test. The remaining datasets were used to evaluate the model's performance after the model's training. Several assessment metrics can be used to carry out the performance evaluation process. This study used four capacity valuation gauging metrics, i.e., MAE (mean absolute error), MSE (mean square error), MAPE (mean absolute percentage error), RMSE (root-mean-square error), and $R^2$(coefficient of determination) that were very helpful in evaluating continuous qualities. To determine the sufficiency and validity of the built prediction models, the calculations of these metrics have been done on testing datasets. The prediction's accuracy is indicated by the $R^2$ coefficient of determination. More accurate predictions are made as a result for $R^2$ that is closer to 1.

In order to understand the likelihood of error in those forecasts, the current study sought to make the most accurate prediction possible. Because they shed light on the possibility of a forecast being inaccurate, these metrics act as superior gauges of accuracy.

#### 6.2.2 Statistical measures

The methodology comprises three steps: constructing machine learning techniques, simulating the malaria model, and formulating a decision matrix based on predicted

output. It concludes with prioritizing machine learning models using the ROVM MCDM technique, ensuring a comprehensive and methodical analysis:

*RMSE* (*Root-mean-square error*) Using a single figure for assessing machine learning model performance, whether in training, or post-deployment monitoring, is highly beneficial [40]. RMSE, a widely used metric, quantifies this. The RMSE equation is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2} \tag{10}$$

where $\left(x_i - \mu\right)^2$ is the square of errors, $N$ stands for the number of mistakes, $x_i$ stands for the observed values, and $\mu$ stands for the predicted values.

*MSE* (*mean square error*) The MSE quantifies the squared discrepancy between model predictions and actual data, averaged across the dataset [40]. The MSE equation is provided below:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 \tag{11}$$

where $\left(x_i - \mu\right)^2$ is the square of errors, $N$ stands for the number of mistakes, $x_i$ stands for the observed values, and $\mu$ stands for the predicted values.

*MAE* (*mean absolute error*) MAE describes inaccuracies between matched observations in data analysis [40]. The Mean Absolute Error equation is provided below:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|z_i - x_i\right| \tag{12}$$

where $z_i$ represents the actual value and $x_i$ represents the predicted value.

*MAPE* (*mean absolute percentage error*): The MAPE [40] is a relative metric that, in essence, scales the mean absolute deviation to be in percentage units rather than the variable's units. The equation of MAPE is given in the Eq. (13):

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_t - Q_t}{P_t}\right| \tag{13}$$

here $P_t$ denotes the actual value and $Q_t$ denotes the predicted value.

$R^2$ R-squared is an indicator of statistical significance that illustrates the proportion of a dependent variable's fluctuation that can be attributed to an independent variable in a regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{\sum_{i=1}^{N}\left(x_i - z_i\right)^2} \tag{14}$$

where $\left(x_i - \mu\right)^2$ is the square of errors, $N$ stands for the number of mistakes, $x_i$ stands for the observed values, and $\mu$ stands for the predicted values. Here $z_i$ denotes the actual value.

### 6.3 Decision matrix

The grading criteria for various regression approaches are presented in this section. The decision matrix is essential to machine learning because it provides an organized way to evaluate and contrast different models or algorithms according to several criteria. This methodical approach promotes optimization and progress in machine learning projects by improving accountability, transparency, and eventually leading to decisions that are well-informed.

#### 6.3.1 Importance of decision matrix

The decision matrix is a fundamental instrument in the complex field of decision-making, valued for its capacity to establish transparency, coherence, and order in complex situations. Within the proposed research, the use of a decision matrix is required due to the need to systematically organize options and standards. When it comes to aligning the decision-making process with the main goal of enhancing machine learning regression algorithms, such careful structure is invaluable.

Using a decision matrix in decision-making procedures is essential for giving complicated situations structure, clarity, and transparency. To optimize machine learning regression methodologies, a decision matrix must be used in the proposed study to systematically arrange criteria and alternatives. Decision-makers may examine options systematically, impartially, and unbiased, based on this organized methodology, which helps them to identify strengths and shortcomings. Examining six statistical metrics using nine various regression procedures, the study shows that performance varies based on several parameters. Decision matrices prioritize possibilities in line with corporate goals, which help with resource allocation as well. The goal of the research is to guarantee optimal results and expedite decision-making processes by creating a decision matrix using regression techniques as alternatives and statistical measurements as criterion.

#### 6.3.2 Formation of decision matrix

In this section, the study provides a brief overview of the process involved in constructing the decision matrix. It has been noted that there is a notable consistency in the results obtained from the comparison between actual and predicted datasets across most regression models employed in this study. Consequently, distinguishing the most effective regression technique for malaria prediction among the nine utilized becomes challenging through mere observation. To meticulously analyze each regression technique and determine the optimal outputs, the introduction of the latest ROVM MCDM techniques is deemed necessary. This necessitates the creation of a decision matrix to facilitate the evaluation process.

The decision matrix of present study consists of 9 rows (consisting of 9 alternatives) and 6 columns (consisting of 6 criteria), i.e., a $9 \times 6$ matrix. Nine different regressions techniques choosen as alternatives (from A1 to A9) as A1: MLP, A2: random forest, A3: SVR (support vector regressor), A4: gradient boosting regressor, A5: Bayesian ridge, A6: kernel ridge, A7: XGB regressor, A8: LGBM (light gradient boosting machine) regressor, and A9: linear regression, and six criteria—R, $R^2$, RMSE, MSE, MAPE, and MSRE choosing from C1 to C6—are chosen to determine the best regression techniques. The ranking of various regressions will be determined by these factors. The whole values of R, $R^2$, RMSE, RMSE, MAPE, and MSRE values are displayed in Table 2, which corresponds to the outputs of various logical predictions for the prediction of malaria illnesses.

### 6.4 Optimal model selection by ROVM MCDM techniques

The range of value method (ROVM) technique [8] was first presented by Yakowitz et al. (1993). It is based only on decision-makers' ordinal assessments of the relevance of the criteria. Thus, the ROVM approach can be quite helpful when decision-makers are having trouble deciding on numerical weights.

An MCDM is a tool used in operations research to identify the best options based on their preferable rank when there are only a limited number of decision criteria available. A technique used in the domain of MCDM is the ROVM, which provides a structured framework to assess and prioritize alternatives according to several criteria. With this approach, the range of values for every criterion that is thought to be important to the decision-making process is established. Decision-makers can then assess and contrast options based on how well they perform within these predetermined ranges by giving these criteria scores or weights. In order to help with informed decision-making, ROVM offers an organized framework for studying complicated choice issues. This framework enables a more

**Table 2** Decision matrix of different regression methods

| Alternatives | | Criteria | | | | | |
|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | C5 | C6 |
| | | RMSE | $R$ | $R^2$ | MAE | MAPE | MSRE |
| A1 | MLP | 0.03357 | 0.99189 | 0.97616 | 0.01071 | 5.95674 | 0.02660 |
| A2 | RF | 0.06385 | 0.97588 | 0.95235 | 0.03733 | 9.82424 | 0.01886 |
| A3 | SVR | 0.10029 | 0.92061 | 0.84753 | 0.06178 | 16.04133 | 0.04523 |
| A4 | GBR | 0.05914 | 0.97433 | 0.94933 | 0.02440 | 11.30079 | 0.04624 |
| A5 | BR | 0.09399 | 0.93074 | 0.86628 | 0.05625 | 14.66863 | 0.05544 |
| A6 | KR | 0.09180 | 0.93269 | 0.86991 | 0.05419 | 15.93820 | 0.06336 |
| A7 | XGB | 0.02748 | 0.99417 | 0.98838 | 0.01058 | 9.87380 | 0.06483 |
| A8 | LGBM | 0.12772 | 0.87443 | 0.76463 | 0.06709 | 15.50018 | 0.05041 |
| A9 | LR | 0.08316 | 0.94630 | 0.89549 | 0.04815 | 12.74019 | 0.03131 |

nuanced understanding of the trade-offs involved. The ROVM holds significant importance for decision-making processes for several reasons:

- With ROVM, decision-makers may evaluate many factors at once, guaranteeing a comprehensive assessment of options for well-informed decision-making.
- ROVM enables a thorough evaluation of errors produced by machine learning models by considering the range of values for each criterion.
- The systematic methodology of ROVM improves transparency in error analysis, assisting users in better comprehending and interpreting errors.

To predict the results of best regression methods, the ROVM (range of value method) MCDM techniques have been used. In the following section, the ROVM MCDM techniques have been discussed:

- Step 1 Employing the range of value methods (ROVM) technique, the study evaluates every single choice's best and worst utility. With this approach, readers may assess the range of possible outcomes for every choice they make in-depth. The research can make better decisions by evaluating both the upper and lower boundaries of utility, which helps us see the possible advantages and disadvantages of each option.
- Step 2 By maximizing and minimizing a utility function, this is accomplished. The following equations are used to determine the best utility ($p_i^+$) and the worst utility ($p_i^-$) for the i$^{th}$ choice for a linear additive model:

$$\max p_i^+ = \sum_{j=1}^{k} p_{ij} q_j \tag{15}$$

$$\min p_i^- = \sum_{j=1}^{k} p_{ij} q_j \tag{16}$$

- Step 3 Regardless of the exact quantitative weights, criteria perform better than alternative k if $p_i^- > p_i^+$. The midpoint, which can be determined as follows, can be used to score the alternatives if it is not possible to distinguish them on this basis (allowing subsequent ranking to be done):

$$p = \frac{p_i^+ + p_i^-}{2} \tag{17}$$

- Step 4 At this final step, all of the alternatives are ranked based on their $p_i$ values. The best alternative is thus shown by the greatest $p_i$ value, while the least desirable alternative is indicated by the lowest $p_i$ value.

This method gives preference to choices that have higher levels of assurance about their results. In order to ensure decision-making based on thorough

statistical analysis, one has to maximize the dependability and trustworthiness of the selected course of action by choosing the option with the greatest $p_i$-value.

## 7 Results and discussion

The proposed methodology elucidates the results of malaria test case prediction. It showcases the machine learning outcomes, MCDM ranking results, and the top three machine learning models. Additionally, a comparison of several MCDM methods ensures the selection of the top machine learning model across all MCDM techniques. This comprehensive approach not only presents the predictive performance of the machine learning models but also validates their effectiveness through rigorous evaluation and comparison using various decision-making methodologies.

### 7.1 Machine learning results

In this section, we will talk about the anticipated result. Eighty percent of the data from the carefully curate collection of sixty datasets were used for training the nine regression models, while the remaining twenty percent are utilized to test the nine specified regression techniques. The results acquired by using different regression method are shown in Table 1. It can be seen that, for each of the nine machine learning algorithms—multilayer perceptron (MLP), random forest (RF), support vector regressor (SVR), gradient boosting regressor (GBR), Bayesian ridge (BR), kernel ridge (KR), extreme gradient boost (XGB) regressor, light gradient boosting

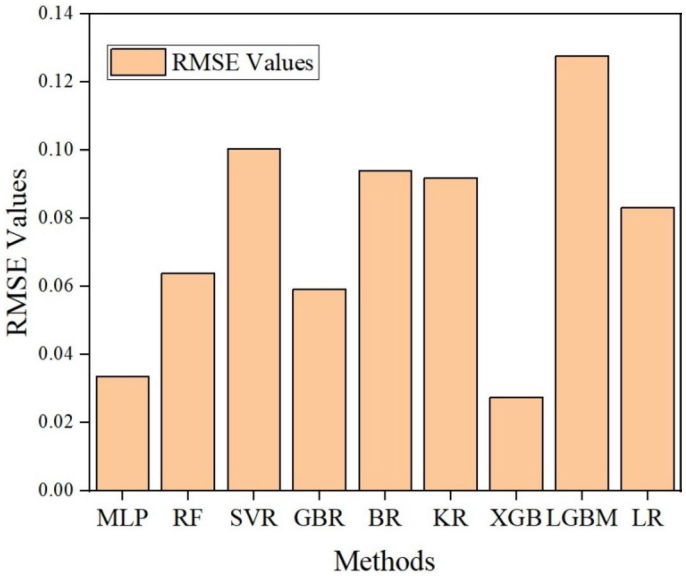


**Fig. 11** RMSE values for each of the nine regression techniques

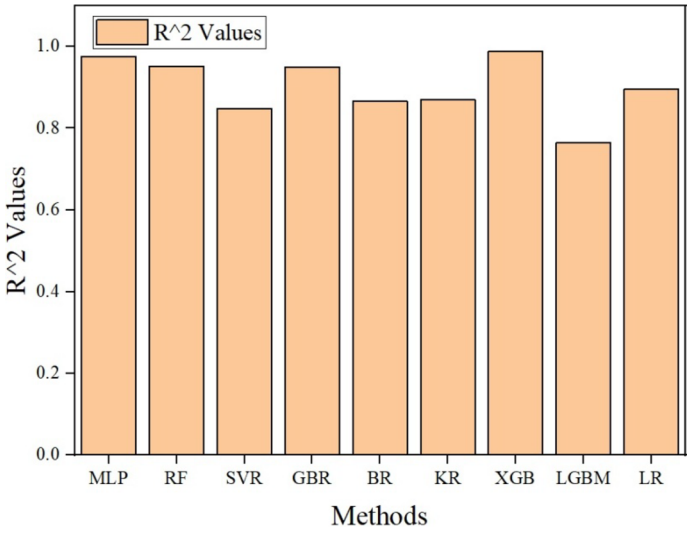


**Fig. 12** $R^2$ values for each of the nine regression techniques

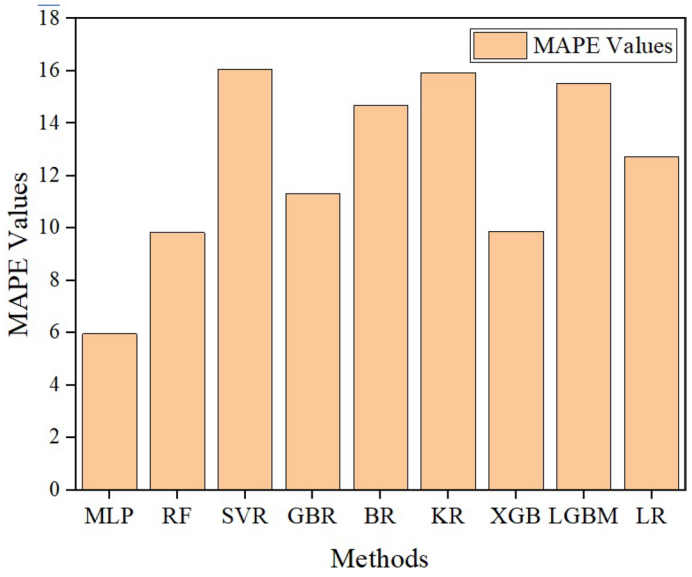


**Fig. 13** MAPE values for each of the nine regressiontechniques

machine (LGBM) regressor, and linear regression (LR)—we get the different R, $R^2$, mean square error, root-mean-square error, mean absolute error, and mean absolute percentage error.

It has been founded that the R value is maximum for the case of XGB regressor and that of $R^2$ is for same regression model, i.e., XGB regressor. The

root-mean-square error (RMSE) is founded minimum for XGB, and the value of RMSE for XGB is 0.02748. The MAPE value is least for MLP regression with a value of 5.95674. The MSRE value is least for random forest regression method with a value of 0. 01886. The graphical representation of the RMSE, $R^2$, and MAPE outcomes for each of the nine regressions employed in this investigation is shown in Figs. 11, 12, and 13.

### 7.2 ROVM results

Multi-criteria decision-making (MCDM) strategies were used in the study to identify the best effective regression models for malaria prediction. The MCDM framework specifically made use of the range of value method (ROVM). Using a variety of criteria, including robustness, interpretability, and accuracy, this approach makes it possible to thoroughly assess different regression algorithms. In Sect. 6.4 of the study, the ROVM-MCDM approach is discussed in detail, with an emphasis on its utility and importance in determining the optimal regression techniques for malaria forecasting. Researchers want to improve the accuracy and dependability of malaria illness prediction by using this method to refine prediction models, which will lead to more successful disease control tactics.

With a $p$ value of 2.84744, the multilayer perceptron (MLP) regression showed the greatest value. Closely after, the XGBoost regressor secured the second-highest rank with a p-value of 2.30578. Higher $p$ values imply greater prediction powers. These values represent the statistical significance of each regression model. Table 3 displays the decision table produced by the ROVM for rank prediction. The results of using the ROVM technique to forecast the ranking of several models are summarized in this table, which offers crucial details about the models' relative effectiveness and fit for the task at issue.

### 7.3 Optimal ML outputs

In this section, the ideal machine learning results for illness prediction will be thoroughly covered. Out of the nine regressions applied in this study, top 3 regressions

**Table 3** Ranking table of ROVM for finding the optimal model

| Methods | $p^+$ | $p^-$ | $p$ | Rank |
|---|---|---|---|---|
| MLP | 1.92631 | 3.76856 | 2.84744 | 1 |
| RF | 1.68623 | 2.78034 | 2.23329 | 3 |
| SVR | 0.75625 | 0.79383 | 0.77504 | 8 |
| GBR | 1.65980 | 2.31396 | 1.98688 | 4 |
| BR | 0.92457 | 0.86871 | 0.89664 | 6 |
| KR | 0.95706 | 0.62869 | 0.79287 | 7 |
| XGB | 2 | 2.61157 | 2.30578 | 2 |
| LGBM | 0 | 0.36725 | 0.18362 | 9 |
| LR | 1.18508 | 1.83617 | 1.51063 | 5 |

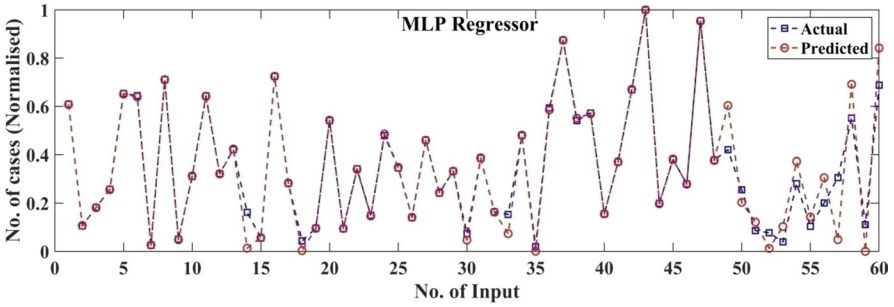


**Fig. 14** Recorded vs. Prediction plot for MLP Regressor

have been chosen to predict the malaria diseases in Tripura. The TOP 3 regressions are subsequently MLP, XGB and random forest. The ranking of these three regressions has been done with the help of ROVM MCDM techniques.

The multilayer perceptron (MLP) regression model's fitness curve, which shows the relationship between actual values and predictions values, is shown in Fig. 14. Across a large range of data points, the curve notably shows a close alignment between the predicted and actual outputs. This synchronization shows that the underlying patterns in the dataset are correctly captured by the MLP regression model. Furthermore, the model's efficacy is further supported by the RMSE and MAE metrics, which show negligible differences between actual and predicted values and have respective values of 0.03357 and 0.01071. These results confirm the MLP regression model's usefulness in the field of predictive modeling by highlighting its strong performance in generating precise predictions.

The fitness curve for the alignment of actual and predicted results using the XGB regressor is shown in Fig. 15. There is almost a perfect match between the observed and predicted instances of malaria due to the model's robust testing score. Remarkably, the XGB regressor attains low RMSE (0.02748) and MAE (0.01058) values, demonstrating its remarkable precision in forecasting the onset of diseases. These results highlight how well the XGB regressor predicts malaria prevalence, which makes it a useful tool for predictive modeling in this situation.

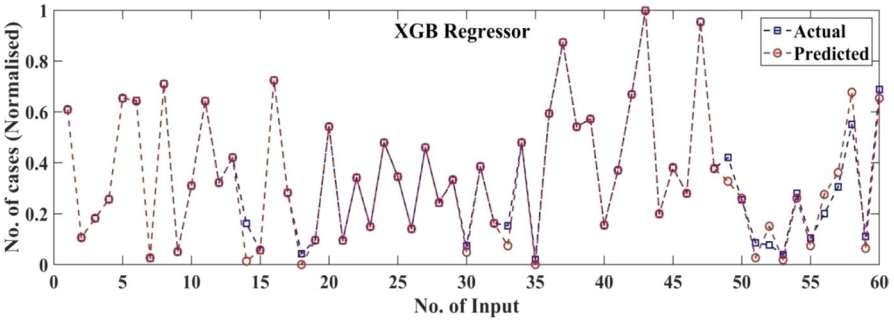


**Fig. 15** Recorded versus Prediction plot for XGB Regressor

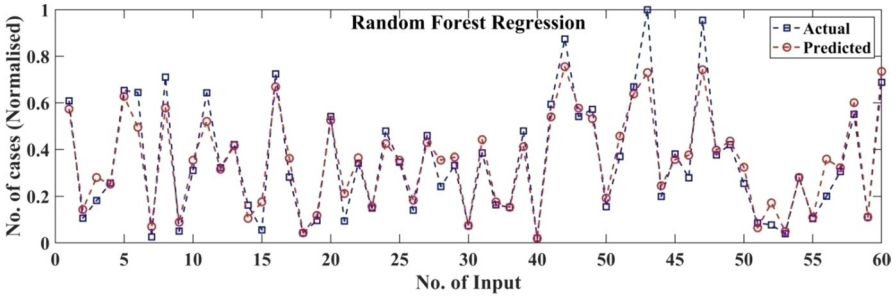


**Fig. 16** Recorded vs. prediction plot for random forest regression

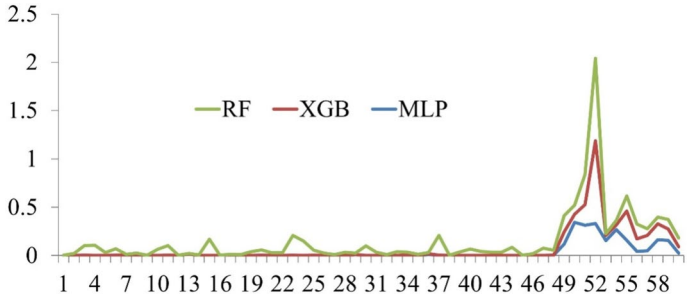


**Fig. 17** Absolute errors of TOP 3 regression techniques found using ROVM approaches

Figure 16 represents the fitness curve of the actual and predicted cases of malaria diseases using random forest regression. It has been observed that the RMSE and MAE values of random forest are quite high than MLP regression and XGB Regressor. The RMSE and MAE value of random forest is, respectively, 0.06385 and 0.03733. The MLP, XGB regressor and the random forest methods are found to be extremely optimal regression methods in predicting the number of malaria cases.

In order to forecast malaria illness, it is now essential to carefully examine the absolute errors between actual and predicted data for the top three regression algorithms found using ROVM approaches. The correlation between the absolute errors of the top three regression models, as determined using ROVM methods, is depicted in Fig. 17. The most appropriate regression technique for the job may be chosen with the help of this study, which offers insights into the precision and dependability of each model in forecasting malaria cases.

### 7.4 Comparison analysis

For several causes, it is essential to compare distinct MCDM approaches when ranking the alternatives. First, it makes performance evaluation easier by enabling decision-makers to evaluate how well each strategy works to solve certain decision-making issues. Second, it additionally makes possible to

**Table 4** Comparison of ROVM techniques with other MCDM techniques for ranking of the alternatives

| | Alternatives | Methods | | | | |
|---|---|---|---|---|---|---|
| | | TOPSIS | SAW | WSM | WASPAS | ROVM |
| A1 | MLP | 1 | 1 | 1 | 1 | 1 |
| A2 | RF | 2 | 2 | 3 | 2 | 3 |
| A3 | SVR | 7 | 7 | 7 | 7 | 8 |
| A4 | GBR | 4 | 4 | 4 | 4 | 4 |
| A5 | BR | 6 | 6 | 6 | 6 | 6 |
| A6 | KR | 8 | 8 | 8 | 8 | 7 |
| A7 | XGB | 3 | 3 | 2 | 3 | 2 |
| A8 | LGBM | 9 | 9 | 9 | 9 | 9 |
| A9 | LR | 5 | 5 | 5 | 5 | 5 |

evaluate robustness under many circumstances, such different criterion weights or unknown data, which sheds light on the reliability and dependability of every method.

Nine distinct regression techniques were ranked in the present study using the ROVM methodologies. A comparison analysis using four additional multiple criteria decision-making (MCDM) approaches was carried out in order to assess the accuracy of the ROVM procedures. Technique for order of preference by similarity to ideal solution (TOPSIS) [41], simple additive weighting (SAW) [42], weighted sum model (WSM) [42], and weighted aggregated sum product assessment (WASPAS) [42] were the techniques used for this comparative analysis.

Table 4 in the study displays nine options sorted based on rankings derived from five different MCDM techniques, including ROVM, TOPSIS, SAW, WSM, and WASPAS. This table offers a detailed comparison of the rankings generated by each method, enabling a full assessment of how well they work to determine which regression approaches are most suited for predicting malaria fever. From this table, it is seen that the XGB, random forest and MLP got the top three ranks based on the most of the MCDM techniques. The alternative XGB regression methodology consistently obtained the second rank across several assessment techniques in the suggested study. In particular, the results of the weighted sum model (WSM) and ROVM approaches showed that XGB regression was the second most successful strategy for predicting malaria fever. This result implies that the suggested study has a noteworthy capacity to forecast the best machine learning regression methods for predicting malaria fever. The consistent top ranking of RF, XGB, and MLP regression underscores their reliability and efficiency as a viable solution for this malaria test cases prediction task.

Figure 18 shows the graph of ranking of the nine alternatives by using five different MCDM techniques. It has been observed that the majority of the alternatives are giving the same ranking by using all the MCDM techniques.

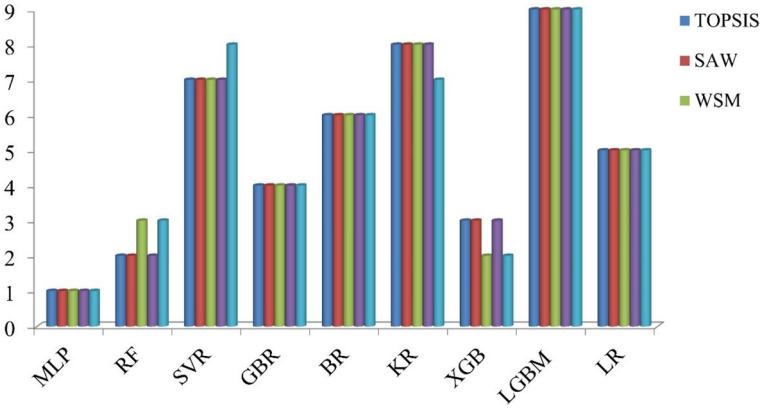


**Fig. 18** Ranking of alternatives by using five MCDM techniques for the comparative analysis

### 7.5 Impact of different feature combination

In this study, we have focused on utilizing various features such as months, year, temperature, precipitation, and humidity to predict malaria cases in Tripura. This section delves into how the combination of these features influences the outcomes of different machine learning models. To conduct this analysis comprehensively, we have explored all possible combinations, including:

- Combination 1 (C1): Temperature, Months, Year
- Combination 2 (C2): Humidity, Month, Year
- Combination 3 (C3): Precipitation, Month, Year
- Combination 4 (C4): Temperature, Humidity, Months, Year
- Combination 5 (C5): Temperature, Precipitation, Months, Year
- Combination 6 (C6): Precipitation, Humidity, Months, Year
- Combination 7 (C7): Proposed Study

For each of these seven combinations, we fitted nine different machine learning (ML) models and meticulously analysed the final outcomes, particularly focusing on the R (correlation coefficient) and RMSE (root-mean-square error) values. Our analysis aimed to discern the effectiveness of each combination in accurately predicting malaria cases in Tripura across a range of ML models. Table 5 meticulously outlines the results of performance evaluations for a range of machine learning algorithms, specifically highlighting how each one fares in relation to distinct combinations of features. Remarkably, our findings indicate that the proposed combination of features consistently outperforms other combinations across all nine ML models. This suggests that the specific combination of months, year, temperature, precipitation, and humidity yields the most reliable and accurate predictions for malaria incidence in Tripura, regardless of the ML algorithm employed.

These results underscore the significance of feature selection and combination in enhancing the predictive capabilities of ML models for tackling malaria outbreaks.

**Table 5** Performance outcomes for different ML models with respect to different combinations

| ML Models | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LR | | LGBM | | KR | | XGB | | BR | |
| | R | RMSE | R | RMSE | R | RMSE | R | RMSE | R | RMSE |
| C1 | 0.88203 | 0.07091 | 0.04419 | 0.20185 | 0.67946 | 0.11689 | 0.65977 | 0.12043 | 0.84799 | 0.08050 |
| C2 | 0.62756 | 0.12736 | 0.71252 | 0.11189 | 0.60941 | 0.13042 | 0.80686 | 0.09171 | 0.62244 | 0.12823 |
| C3 | 0.72959 | 0.12304 | 0.45099 | 0.17532 | 0.77134 | 0.11315 | 0.83110 | 0.00946 | 0.66885 | 0.13616 |
| C4 | 0.72256 | 0.15802 | 0.06938 | 0.28941 | 0.61420 | 0.18634 | 0.79470 | 0.13593 | 0.71161 | 0.16111 |
| C5 | 0.60701 | 0.16470 | 0.50315 | 0.18519 | 0.49287 | 0.18709 | 0.55304 | 0.14322 | 0.68276 | 0.12066 |
| C6 | 0.64093 | 0.14560 | 0.63620 | 0.16726 | 0.43938 | 0.15151 | 0.54242 | 0.16436 | 0.60432 | 0.15284 |
| C7 | 0.94630 | 0.08316 | 0.87443 | 0.12772 | 0.93269 | 0.09180 | 0.99417 | 0.02748 | 0.93074 | 0.09399 |

| ML Models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | GBR | | SVR | | RF | | MLP | |
| | *R* | RMSE | *R* | RMSE | *R* | RMSE | *R* | RMSE |
| C1 | 0.49314 | 0.14699 | 0.85808 | 0.07778 | 0.84660 | 0.08086 | 0.87135 | 0.07406 |
| C2 | 0.74082 | 0.10624 | 0.80885 | 0.09124 | 0.87316 | 0.07432 | 0.97964 | 0.25062 |
| C3 | 0.80833 | 0.10359 | 0.91746 | 0.06798 | 0.90124 | 0.07436 | 0.69470 | 0.13074 |
| C4 | 0.62510 | 0.18369 | 0.60392 | 0.18881 | 0.69104 | 0.16676 | 0.55326 | 0.20052 |
| C5 | 0.53828 | 0.14556 | 0.56982 | 0.14050 | 0.59016 | 0.13714 | 0.55864 | 0.17454 |
| C6 | 0.53064 | 0.16646 | 0.68150 | 0.13713 | 0.57592 | 0.15823 | 0.99937 | 0.16726 |
| C7 | 0.97433 | 0.05914 | 0.92061 | 0.10029 | 0.97588 | 0.06385 | 0.99189 | 0.03357 |

By leveraging the optimal combination of features, we can significantly improve the accuracy and efficacy of malaria prediction models, thus facilitating more informed decision-making and better resource allocation for disease prevention and control efforts in Tripura. By systematically examining these combinations, we aim to understand their collective impact on the predictive capabilities of our models. This analysis provides valuable insights into which combinations yield the most accurate results, contributing to the refinement and optimization of our predictive methodologies for combating malaria in Tripura.

## 8 Conclusion and future scope

India, a vast country, has a billion people who could contract malaria. Due to the diverse habitat, accessibility of infectious disease vectors, and biological traits, disease transmission is challenging. It is estimated that 4% of India's population lives in the northeast, where malaria is common and consistently contributes 10% of cases (the bulk of which are Plasmodium falciparum each year) and 20% of reported deaths. Therefore, it has become very difficult to lower the danger of malaria. If we can predict the malaria cases, then it becomes easy to diagnose the diseases.

This paper provides a thorough overview of how methods from machine learning are used to forecast desired results. As the incidence of malaria rises with rising temperatures, humidity levels, and precipitation, the variables have been brought into account as inputs for the current study. We can gain information from the current study about how to pick the best approach to malaria illness prediction.

Nine different regression techniques, including multilayer perceptron (MLP), random forest (RF), support vector regressor (SVR), gradient boosting regressor (GBR), Bayesian ridge (BR), kernel ridge (KR), extreme gradient boost (XGB) regressor, light gradient boosting machine (LGBM) regressor, and linear regression (LR), were used in the current study. In the current investigation, R, $R^2$, RMSE, MAE, MAPE, and MSRE were utilised as indicators of statistical significance and considering these measures the decision matrix for the MCDM techniques has been formed. From Table 1, it has been observed that the minimum RMSE value is obtained by MLP regression with a value of 0.03357. Also, the MAE value for the MLP regression is very as compare to other models. To select the most appropriate model, ROVM (range of value method) MCDM techniques has been used. From graph 9, the actual and anticipated numbers of malaria cases, as determined by MLP regression algorithms, nearly always coincide. The best regression technique with the highest degree of accuracy, it follows, is MLP regression. Finally, various MCDM methods such as WSM, WASPAS, and WPM were utilized to prioritize the applied machine learning (ML) model for this specific application. Through this process, it was demonstrated that the random forest (RF), multilayer perceptron (MLP), and XGBoost (XGB) emerged as the top three ML models across the majority of the MCDM cases. This rigorous evaluation ensures that the selected ML models are not only effective but also well suited for addressing the challenges and requirements of the given application, providing confidence in their reliability and performance. It could be helpful for governmental, profit-making, and nonprofit organizations to continue their intervention strategies in order to successfully combat the disease in Tripura.

The study's limitations include the proposed ML regression technique shows promise; its real-world effectiveness remains to be fully validated through prospective studies. Also, the ML model can be improved through several ensemble techniques which could be our future goal of the study. The focus of the current study was on machine learning methods for predicting malaria cases in the state of Tripura. Our long-term goal is to research how malaria will spread throughout India. In order to calculate outcomes more precisely, new methodologies can be introduced. Although utilized just five variables in this model, more inputs can be included in the future to examine diseases more thoroughly.

**Acknowledgements** The authors are appreciative for the Malaria Data assistance provided by the Chief Medical Officer, District Hospital, Dharmanagar, North Tripura, Tripura, India. The authors are appreciative for the assistance provided by the National Board of Higher Mathematics (NBHM) under the Department of Atomic Energy (DAE), India, for the support of the research study at the Department of Mathematics, NIT Agartala.

**Author contributions** Apurba Debnath contributed to written-original draft, Anirban Tarafdar contributed to conceptualization, validation, and formal analysis. A Poojitha Reddy and Azharuddin Shaikh

contributed to software and figures preparation. Paritosh Bhattacharya contributed to resources and supervision.

**Funding** There is no funding for this research.

**Data availability** Given in the manuscript. The code is available in the following link: https://github.com/Apurba1240/MultiTimeModeling/blob/main/Copy_of_Filtration.ipynb

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Ethical approval** Not applicable.

## References

1. Vector-borne diseases (2020) https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases. Accessed 12 May 2022
2. Gomez-Elipe A, Otero A, Van Herp M, Aguirre-Jaime A (2007) Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997–2003. Malar J 6(1):1–10. https://doi.org/10.1186/1475-2875-6-129
3. Thomson MC, Mason SJ, Phindela T, Connor SJ (2005) Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. https://doi.org/10.7916/d8-v495-fa43
4. World Health Organization. World malaria report 2020: 20 years of global progress and challenges. World Health Organization (2020) https://www.who.int/news-room/questions-and-answers/item/malaria?gclid=CjwKCAjwx_eiBhBGEiwA15gLN__D428166i1MW8NchKbJljcHufDn-kFQcsVN1Gc_sbM3hBSHaPnchoCVyQQAvD_BwE. Accessed 12 May 2022
5. Liu Z, Wang S, Zhang Y, Xiang J, Tong MX, Gao Q, Bi P (2021) Effect of temperature and its interactions with relative humidity and rainfall on malaria in a temperate city Suzhou, China. Environ Sci Pollut Res 28:16830–16842. https://doi.org/10.1007/s11356-020-12138-4
6. Kumar P, Vatsa R, Sarthi PP, Kumar M, Gangare V (2020) Modeling an association between malaria cases and climate variables for Keonjhar district of Odisha, India: a Bayesian approach. J Parasit Dis 44:319–331. https://doi.org/10.1007/s12639-020-01210-y
7. Statista, Number of malaria case across India (2022) https://www.statista.com/statistics/976130/number-of-malaria-cases-india/. Accessed Dec 2022
8. Madić M, Radovanović M, Manić M (2016) Application of the ROV method for the selection of cutting fluids. Decis Sci Lett 5(2):245–254. https://doi.org/10.5267/j.dsl.2015.12.001
9. Lee YW, Choi JW, Shin EH (2021) Machine learning model for predicting malaria using clinical information. Comput Biol Med 129:104151. https://doi.org/10.1016/j.compbiomed.2020.104151
10. Santosh T, Ramesh D, Reddy D (2020) LSTM based prediction of malaria abundances using big data. Comput Biol Med 124:103859. https://doi.org/10.1016/j.compbiomed.2020.103859
11. Kim Y, Ratnam JV, Morioka Y, Behera S, Tsuzuki A, Minakawa N, Hashizume M (2019) Malaria predictions based on seasonal climate forecasts in South Africa: a time series distributed lag nonlinear model. Sci Rep 9(1):1–10. https://doi.org/10.1038/s41598-019-43372-7
12. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19(1):1–16. https://doi.org/10.1186/s12911-019-1004-8
13. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE access 7:81542–81554
14. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad AKM, Uddin S, Kamal AHM, Moni MA (2021) Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. JMIR Med Inform 9(4):e25884. https://doi.org/10.2196/25884

15. Nezu N, Usui Y, Saito A, Shimizu H, Asakage M, Yamakawa N, Goto H (2021) Machine learning approach for intraocular disease prediction based on aqueous humor immune mediator profiles. Ophthalmology 128(8):1197–1208. https://doi.org/10.1016/j.ophtha.2021.01.019
16. Africa CW, Turton M (2019) Oral health status and treatment needs of pregnant women attending antenatal clinics in KwaZulu-Natal, South Africa. Int J Dent. https://doi.org/10.1155/2019/5475973
17. Rachmi CN, Agho KE, Li M, Baur LA (2016) Stunting, underweight and overweight in children aged 2.0–4.9 years in indonesia prevalence trends and associated risk factors. PLoS ONE 11(5):e0154756. https://doi.org/10.1371/journal.pone.0154756
18. World Health Organization (WHO) (2015) https://www.who.int/teams/global-malaria-programme. Accessed 12 May 2022
19. Alanazi R (2022) Identification and prediction of chronic diseases using machine learning approach. J Healthc Eng. https://doi.org/10.1155/2022/2826127
20. Nkiruka O, Prasad R, Clement O (2021) Prediction of malaria incidence using climate variability and machine learning. Inform Med Unlocked 22:100508. https://doi.org/10.1016/j.imu.2020.100508
21. Tarafdar A, Majumder P, Bera UK (2023) Prediction of air quality index in kolkata city using an advanced learned interval type-3 fuzzy logic system. In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), pp 1–7
22. Debnath R, Majumder P, Tarafdar A, Bhattacharya B, Bera UK (2024) Artificial intelligence based supply chain management strategy during COVID-19 situation. In: Supply chain forum. Taylor & Francis, pp. 1–20. https://doi.org/10.1080/16258312.2024.2303307
23. Debnath A, Tarafdar A, Bhattacharya P, Shaikh A (2023) MOORA MCDM based optimal machine learning regression techniques for breast cancer prediction. In: 2023 IEEE Silchar Subsection Conference (SILCON), pp 1–7
24. Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, Tariq U (2021) An AI-based intelligent system for healthcare analysis using ridge-adaline stochastic gradient descent classifier. J Supercomput 77:1998–2017. https://doi.org/10.1007/s11227-020-03347-2
25. Veeramakali T, Siva R, Sivakumar B, Senthil Mahesh PC, Krishnaraj N (2021) An intelligent internet of things-based secure healthcare framework using blockchain technology with an optimal deep learning model. J Supercomput 77(9):9576–9596. https://doi.org/10.1007/s11227-021-03637-3
26. Weather and Climate (2018) https://tcktcktck.org/india/tripura/dharmanagar/december-2018. Accessed 12 May 2022
27. Past Weather in Agartala, March (2021) https://www.timeanddate.com/weather/india/agartala/historic?month=3&year=2021. Accessed 10 April 2023
28. Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, Hoek G (2019) A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ Int 130:104934. https://doi.org/10.1016/j.envint.2019.104934
29. Taki M, Rohani A, Soheili-Fard F, Abdeshahi A (2018) Assessment of energy consumption and modeling of output energy for wheat production by neural network (MLP and RBF) and Gaussian process regression (GPR) models. J Clean Prod 172:3028–3041. https://doi.org/10.1016/j.jclepro.2017.11.107
30. Guo Z, Yu B, Hao M, Wang W, Jiang Y, Zong F (2021) A novel hybrid method for flight departure delay prediction using random forest regression and maximal information coefficient. Aerosp Sci Technol 116:106822. https://doi.org/10.1016/j.ast.2021.106822
31. Poojitha Reddy A, Tarafdar A, Kumar Bera U (2023) Regression based machine learning approach to predict flight price between Bangalore and Kolkata. In: IEEE 8th International Conference for Convergence in Technology (I2CT) Pune, India
32. Keprate A, Ratnayake RMC (2017) Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping. In: IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, pp 1331–1336. https://doi.org/10.1109/IEEM.2017.8290109
33. Shi Q, Abdel-Aty M, Lee J (2016) A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accid Anal Prev 88:124–137. https://doi.org/10.1016/j.aap.2015.12.001
34. Saqib M (2021) Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. Appl Intell 51(5):2703–2713. https://doi.org/10.1007/s10489-020-01942-7
35. Zhang Y, Duchi J, Wainwright M (2013) Divide and conquer kernel ridge regression. In: Proceedings of the 26th Annual Conference on Learning Theory, in proceedings of machine learning research. 30: 592–617, Available from https://proceedings.mlr.press/v30/Zhang13.html

36. Shehadeh A, Alshboul O, Al Mamlook RE, Hamedat O (2021) Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, lightgbm, and xgboost regression. Autom Constr 129:103827. https://doi.org/10.1016/j.autcon.2021.103827
37. Poole MA, O'Farrell PN (1971) The assumptions of the linear regression model. Trans Inst Br 52:145–158. https://doi.org/10.2307/621706
38. Rhodes CL (1996) The process simulation revolution: thermophysical property needs and concerns. J Chem Eng Data 41(5):947–950. https://doi.org/10.1021/je960029b
39. Ketu S (2022) Spatial air quality index and air pollutant concentration prediction using linear regression based recursive feature elimination with random forest regression (RFERF): a case study in India. Nat Hazards 114(2):2109–2138. https://doi.org/10.1007/s11069-022-05463-z
40. Chicco D, Warrens MJ, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7:e623. https://doi.org/10.7717/peerj-cs.623
41. Chakraborty S (2022) TOPSIS and modified topsis: a comparative analysis. Decis Anal 2:100021. https://doi.org/10.1016/j.dajour.2021.100021
42. Chakraborty S (2014) Applications of WASPAS method in manufacturing decision making. Informatica 25(1):1–20. https://doi.org/10.15388/Informatica.2014.01