



# Prediction of seasonal infectious diseases based on hybrid machine learning approach

K. Indhumathi<sup>1</sup> · K. Satheshkumar<sup>2</sup>

Received: 4 January 2023 / Revised: 25 April 2023 / Accepted: 29 May 2023 /  
Published online: 5 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

The effective management of seasonal dengue fever and other viral diseases fever such as malaria, pneumonia, and typhoid fever requires the early deployment of control measures. Predicting disease outbreaks accurately can aid in gaining control of epidemic seasons. In this research, a machine learning-based prediction model is suggested for predicting seasonality diseases. The model uses real-time data collected from different regions around Madurai district between 2019 and 2020, with 29 features including illness such as dengue, malaria, pneumonia, typhoid, kala-azar, Japanese encephalitis, measles, and normal fever and cold infections. The proposed model is a hybrid approach that includes feature selection using the Antlion Optimization Algorithm (ALO) and classification using Random Forest (RF) integrated with the XG-Boost technique. The suggested model's efficiency is examined by accuracy, precision, recall, specificity, and f1-score as performance metrics, and compared with other models such as ACO-ANN, PSO-RF, WO-RF, and ANOVA-SVM. The suggested framework attained a high level of precision of 96.17%, a precision of 93.95%, a recall of 95.86%, a specificity of 93.23%, and an f1-score of 96.22%. Based on the comparison, the suggested model surpassed the efficiency of the alternative methods compared models in terms of all the parameters.

**Keywords** Disease prediction · Hybrid machine learning model · Feature selection · ALO · Random Forest · XG-boost

## 1 Introduction

Malaria, typhoid, pneumonia, and dengue diseases can have an important effect on a person's health and could lead to death if they are not treated properly [9]. These diseases are classified as Airborne, Waterborne, and Vector-Borne. Droplets of

---

✉ K. Indhumathi  
indhu16aug@gmail.com

K. Satheshkumar  
sathesh.drl@gmail.com

<sup>1</sup> Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Srivilliputhur, Tamil Nadu 626126, India

<sup>2</sup> Department of Computer Science and Information Technology, Kalasalingam Academy of Research and Education, Krishnankoil, Srivilliputhur, Tamil Nadu 626126, India

microorganisms that are discharged into the air by cough, sneezing, or talking cause airborne diseases [15]. The pathogens in question may be viruses, bacteria, or fungi. Tuberculosis, influenza, and smallpox are just a few of the diseases that can be transmitted through the air. Pathogenic bacteria are the most prevalent source of waterborne infections, which are transferred primarily through unclean fresh water. Bathing, washing, drinking, and eating food that has been infected can transmit infections. Some of the most serious dangers to human health are Malaria, dengue fever, and West Nile virus are examples of diseases transmitted by mosquitoes. Insects infected with the disease can spread it to humans, as can direct transmission between humans [10].

Weather conditions must be just right for germs to survive, reproduce, and spread. Unpredictable weather conditions have a significant impact on the environment's resources. That's why it's important to focus on changes in the temperature or weather. Viruses like malaria and cholera have been linked to severe rains, for example. This cyclicity is found in acute infectious disorders and may be found in human infectious diseases generally. There is a seasonal window of incidence for each acute infectious disease, and these windows can range significantly among geographic areas and from other diseases within the same region. When epidemics occur, the seasonal fluctuation in infectious disease transmission is an essential factor, but it is not the only factor. Variations in seasons are regarded as "seasonal forcing" in Ecology of infectious diseases and epidemic modelling, respectively [12].

The seasonal drivers have been environmental factors, particularly climate conditions. This may be because seasonal disease incidence tends to follow a similar pattern. Temperature and rainfall are classic examples of abiotic environmental drivers that influence transmissions through their effects on parasites or hosts, but various examples incorporate seasonal non-climatic abiotic environments including the salinity of the water that might influence water-borne pathogens. Pathogen survival might be affected by environmental conditions when moving between hosts. For instance, droplet-transmitted diseases may have short transition windows (e.g., a few minutes) or longer transition windows (like several weeks) (e.g., parasites with life stages of environments). Other than infections, environmental factors can impact host vulnerability to infections or the dynamics of vector populations [11].

In this research, the main objective is to predict seasonal diseases based on the proposed ML model. For the evaluation, a dataset was collected from different regions around the Madurai district in Tamilnadu. Based on the features from the dataset, the ANOVA technique is used for the feature selection, and prediction and a hybrid random forest with the XG-Boost technique is used. A brief introduction about seasonal diseases was presented in Section 1 and the remaining part of this paper was divided into the areas listed below. Section 2 represents a discussion of the related works and their applications, Section 3 represents the overview of the suggested model, Section 4 explains the experimental analysis, and the conclusion is presented in Section 5.

## 1.1 Contribution and significance of the work

The contribution of this research includes proposing a hybrid machine-learning model for predicting seasonal ailments such as dengue fever, malaria, pneumonia, and typhoid are examples of diseases based data from the Madurai district in real time. The model utilizes Antlion Optimization Algorithm for feature selection and Random Forest integrated with the XG-Boost technique for classification, achieving high accuracy, precision, recall, specificity, and f1-score, outperforming other compared models.

## 2 Related works

A group forecasting model based on bragging rights [17] was used to anticipate the amount of illness episodes determined by previous events. Log transformation and z-score transformations were used in the preprocessing. A specific disease dataset was used to demonstrate the applicability of this ensemble model. This model was tested for dengue, tuberculosis, chickenpox, and food poisoning among other infections. In [8], a comparative analysis of machine learning classifiers' performance was presented based on disease prediction. Classifiers such as Naive Bayes and decision models, and RF were utilized in the development of the disease prediction system. The symptoms from the obtained datasets for 41 different diseases were used to evaluate this disease prediction algorithm.

AI approaches were employed in [4] to forecast diseases using a feature selection method. Classifiers such as decision trees, SVM, random forest, and XG-Boost algorithms were utilized for the classification of diseases like Malaria, Jaundice, COVID, Pneumonia, Dengue, and Typhoid. Dengue fever/dengue hemorrhagic fever (DF/DHF) has been a problem in Rajasthan for the past decade, and in [2] a model was proposed to predict the prevalence of DF/DHF in 2011. A SARIMA model was employed to conduct statistical analysis. Research shows that adding a forecasting model to a current disease control program can have a significant influence on reducing disease incidence.

To better understand seasonal influenza, a model created and analyzed an epidemiological model known as Seasonal SIR (Seasonal SIR) [1]. To better understand how vector-borne diseases propagate, this model takes into account temperature fluctuations and a variety of human interaction networks. The Seasonal SIR model has been shown to accurately replicate outbreak dynamics observed in the actual world. More accurate forecasts of disease outbreaks may now be made, allowing for the implementation of effective control measures to keep epidemics under control.

LASSO and RF regressions with LSTM, a DRNN was proposed in [13] and compared to forecast dengue incidence on a weekly basis in 790 Brazilian cities. Predictors were based on multivariate time series, and time series from related regions were employed for capturing the spatial components of disease transmissions. Predicting future dengue outbreaks in towns of varying sizes was made easier with the use of an LSTM RRN model. An LSTM model developed in [7] shown to be successful in predicting epidemic disease-infected areas and intensity.

In [5], a PSO-ANN-based system was proposed. The ANN method optimized using the PSO methodology in this model. Dengue patients were identified using a PSO-optimized ANN method. This diagnostic approach was found to be an effective and powerful tool for detecting dengue disease early and more accurately. A diagnostic model for predicting dengue sickness, developed by the same author in [6] using three widely used machine learning approaches (Decision Tree, Artificial Neural Network, and Naive Bayes), was examined.

In the prediction problem of an infectious epidemic, a new scalable feature selection approach and a seasonal adjustment-based predictive modelling framework were proposed. [18]. More stable and accurate epidemic predictions against both long and short-term variations in search engine logs were accomplished by evaluating each component separately. A machine learning strategy based purely on human genome data was reported in [3] for the predictions of dengue disease severity depends on the genome information of humans. The SVM algorithm was used for finding the classification subset of optimal loci and ANN was utilized for classifying patients into those with mild or serious dengue fever.

Previous research in disease prediction using machine learning has shown promising results in terms of accuracy and efficiency. However, some limitations and knowledge gaps need to be addressed. Some of the benefits and drawbacks of relevant research.

## 2.1 Benefits

Improved disease, Early warning systems, Real-time data utilization.

## 2.2 Drawbacks

Variability in datasets, Limited scalability, Lack of interpretability.

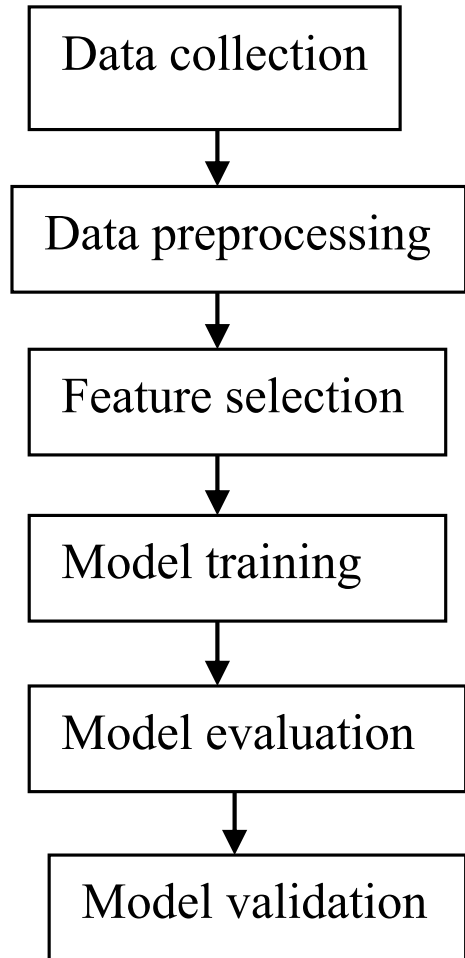
This publication addresses some of the current trends and knowledge gaps in disease prediction research. Specifically, it proposes a hybrid machine learning model that incorporates the Antlion Optimization Algorithm for feature selection and Random Forest integrated with XG-Boost for classification, achieving high accuracy and performance. The research also highlights the rigorous analysis of the dataset selection process to ensure the reliability and representativeness of the real-time data from different regions around the Madurai district. Furthermore, the proposed model has the potential for broader applications beyond the Madurai district and can be extended to incorporate climatic and weather changes for disease prediction. This addresses the current trend of incorporating environmental factors in disease prediction models, considering the impact of climate change on disease dynamics.

## 3 Comprehensive experimental flowchart

The experimental flowchart Fig. 1 provides a comprehensive overview of the proposed strategy for predicting seasonality diseases using a hybrid machine learning model, from data collection to model evaluation and optimization.

## 4 Proposed methodology

The proposed model aims to predict and classify different diseases, including dengue, malaria, pneumonia, typhoid, kala-azar, Japanese encephalitis, measles, and normal fever and cold infections, using real-time data gathered from numerous sources around the Madurai district between 2019 and 2020. The suggested model is a hybrid machine-learning approach that consists of two main steps: feature selection and classification. A critical phase is choosing features in machine learning, as it involves identifying the most relevant features from the dataset that contribute the most to prediction accuracy. In this model, the Antlion Optimization Algorithm (ALO) is utilized for feature selection. ALO is a metaheuristic technique by the hunting behavior of antlions, and it is employed to choose the best qualities from the dataset that are most informative for disease prediction. After feature selection, the Random Forest (RF) algorithm, integrated with the XG-Boost technique, is used for classification. Multiplexing is the technique of combining multiple signals into a single transmission channel to efficiently utilize the available bandwidth. Time-division multiplexing (TDM) involves dividing the transmission channel into discrete time slots, with each time slot dedicated to a different signal. The signals are interleaved in time and transmitted sequentially. Frequency-division multiplexing (FDM) involves dividing the available bandwidth into multiple frequency bands, with each band dedicated to a different signal. The signals are transmitted simultaneously over their respective frequency bands. TDM and FDM are commonly used in telecommunications and networking to

**Fig. 1** Experimental flowchart

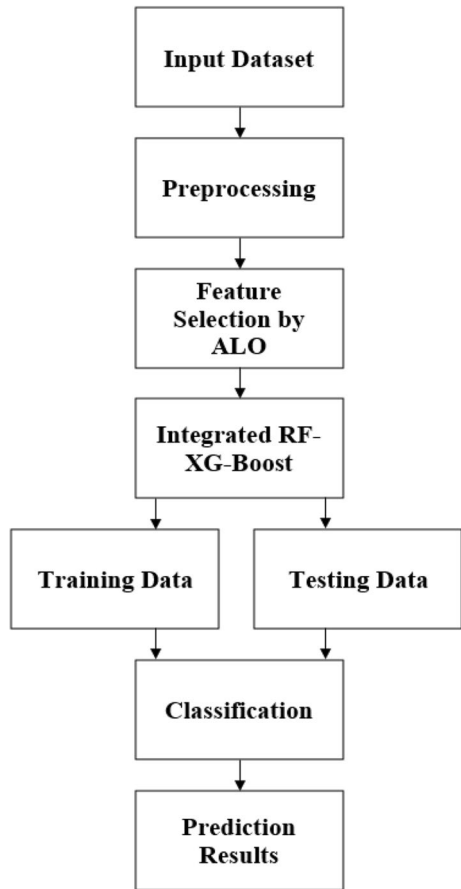
enable multiple signals to be transmitted over a single channel, allowing for efficient use of resources and increased capacity., RF is a robust ensemble learning technique which integrates numerous decision trees and accurate prediction model. XG-Boost is a popular gradient-boosting technique that enhances the performance of decision tree-based models by addressing issues like overfitting and improving prediction accuracy.

Once the classification is done, the results are compared with the original data for validation. This involves evaluating the efficacy of the proposed model in relation to performance metrics, which are commonly used performance metrics in classification tasks. This comparison helps to assess the efficiency of the proposed method in predicting and classifying seasonality diseases based on the live data obtained from the Madurai region in (Fig. 2).

## 5 Feature selection using ALO

As the name implies, this is a method of searching that emulates the hunting strategy used by the antlions in common. ALO has a significant potential to avoid local optima stagnations due to its usage of the roulette wheel and random walk. In ALO,

**Fig. 2** Workflow of the proposed model



the antlion's random selection and the ant's random movement around it assure searching for space, while the adaptative lowering boundaries of antlion traps check that the search region is being used.

The discrete optimization problem that decimal coding cannot directly address is the question of feature selection. For all individuals in the binary arrays, the binary coding version translates the population value of probability into the value of 1 or 0 for all binary vectors. Consequently, all ants in ALO are encoded using the binary coding method. A 0 or a 1 can only represent a discrete binary condition's one dimension. Variables change from 1 to 0 or vice versa when traveling across a dimension. Assume that each ant is a continuous algorithm to implement binary modes for ALO. Binary ALO updates ants by switching between "0" and "1," which is the primary difference between binary and standard ALO. With the probabilities given in Eq. 1, it was assumed that the new ant's location would be either 0 or 1, which was updated by a condition.

$$S_j^k = \begin{cases} 1 & \text{if } rdm < \left| \tanh \left( GZ_j^k \right) \right| \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Hyperbolic tangent functions are represented by  $\tanh$ , while the representation of an ant's location using binary code  $S_j^k$ . Eq. 1 represents a function that generates a binary output (1 or 0) based on a comparison between a random variable  $rdm$  and the absolute value of the hyperbolic tangent of  $GZ_j^k$ , where  $GZ_j^k$  is likely a calculated value based on inputs or parameters. The specific use and interpretation of this function would depend on the context in which it is being used.

Due to elitism, which allows them to keep the best solutions at any stage of optimizations, swarm intelligence algorithms have a substantial advantage over traditional binary coding. Many parent solutions are required for crossover, and the child solutions are generated from across the entire population. It was a comparison of the two binary solutions generated by the random motion. The best antlion was kept as an elite in each iteration. Each ant on the repetition must be influenced by the movement of the strongest and fittest of the antlions. As a result, according to Eq. 2, it was assumed that ants travel between the elite and antlions following the roulette wheel.

$$Ant_j^k = Crossover(S_p^k, S_C^k) \quad (2)$$

Equation (2) represents the crossover operation in the Antlion Optimization Algorithm (ALO), where “Ant” refers to the ants in the algorithm, “j” denotes the ant index, and “k” denotes the iteration number or the current generation of the algorithm.  $S_p^k$  represents the position of the parent ant in the k-th iteration, and  $S_C^k$  represents the position of the child ant generated through crossover in the same iteration.

Wrapper-based approaches require a lot of training and testing for a single assessment of the provided solutions. This means that the method of the search must be carefully chosen. This work used the ALO for searching the feature spaces for a combination of the optimal feature that maximized classification accuracy. Prey positions in the Antlion optimizer are randomly chosen to represent a particular feature combination. Additionally, Antlion tries to search for its prey by setting up n random positions for each of its ants. Using a roulette wheel method, an antlion is chosen for hunting by the antlion method, and randomly walks ants around this ant, as well as randomly walks the best/elite antlions. A given ant adjusts its location based on its last two random treks. If the fitness level of an ant exceeds that of an antlion, then the antlion consumes it and moves into the ant's place. This procedure is applied iteratively in the hopes of reaching a satisfactory conclusion after some iterations. It's worth noting that the ant's random walk is currently constrained by its exploration rate. To facilitate extensive search, it is usual practice to reduce the exploration rate as optimization continues. As many qualities as there are in the supplied data collection, a continuous-valued vector represents each solution. It is restricted to a range of [0,1] for the solution vector. The binary form of approach with a continuous value is the threshold during the solution fitness function evaluation using Eq. (3).

$$y_{ij} = \begin{cases} 0 & \text{if } (x_{ij} < 0.5) \\ 1 & \text{Otherwise} \end{cases} \quad (3)$$

Here,  $x_{ij}$ ,  $i$  is the solution in dimension  $j$ , and  $y_{ij}$  is the solution vector  $x$  in discrete representation [19].

*Input: Fitness function, Search space, Number of Antlions, ants, and iterations (T)*

*Output: The Elitist Antlion and the its fitness*

1) *Set the population of  $n$  antlions and  $n$  ant's positions randomly.*

2) *Compute the fitness of each antlion and ant.*

3) *Discover the fittest antlion; Elite.*

4)  $t=0$ .

5) *while ( $t \leq T$ )*

*for each anti do*

- *Choose the antlion using Roulette wheel (building trap).*

- *Slide ants near the antlion;*

- *Create a random walk for this anti and normalize it;*

*end*

6) *Compute the fitness of each ant.*

7) *Change the antlion with its related ant if it becomes fit (Catching Prey);*

8) *Update elite if the antlion becomes fit than elite.*

*end while*

## 6 Integration of random Forest with XG-boost

Regarding classifications and regressions tree (CART), the random forest was a strong ensembled learning approach. Regression and classification can be applied to it because it has been widely utilized and has a good performance. The theory of statistical learning called bootstraps resampling is used for extracting several samples from the real sample, modeling decision trees for every bootstrapped sample, and hence integrating the prediction of several decision trees for averaging the final prediction outputs. Put-back samples and random changes in the integration of predictors are used to boost the diversity of decision trees. Regression trees (RT) could be useful for this work, which was to predict regressions. It was determined at every branching point of RT that the mean square error between samples was calculated for each leaf node. The regression tree will stop growing if the minimum mean square error at each leaf node is used as the branching condition. This can be done until there are no more characteristics were present or the total mean square error was optimal. The count of regression trees ( $N$  estimators) and the count of the node's random variables are two important custom factors (Max depth). To reduce the number of errors that takes place during data processing, these parameters must be improved.

The RF algorithm's modeling approach is as follows: extract  $N$  from the initial data collection estimations and training sets are generated using bootstrap sampling. Approximately two-thirds of the original information set is used for training. For each random forest training process bootstrap sample, only around a third will be picked. Out-of-bag data is the name given to this portion of the data.

To create a "forest,"  $N$  estimators' total regression trees were built, but these trees are not pruned. No attribute is picked as an internal node for branching in each tree, but a random Max depth attribute is chosen from a pool of these characteristics to serve as a branching node. Thus, the RF approach improves the combined regression analysis



model’s extended forecasting capability by increasing the difference between the regression models by building various training sets. To build multi-RM designs, a RM of  $\{t_1(x), t_2(x), \dots, t_k(x)\}$  was created by n-time model training. Then, using the regression tree predictions of the  $N$  estimators, determine the new sample’s value using a simple average technique. The formula for regression decision is,

$$\hat{f}_{rf}^K(x) = \frac{1}{K} \sum_{k=1}^K t_i(x) \tag{4}$$

here  $\hat{f}_{rf}^K(x)$  denotes the model of mixed regression,  $t_i$  is a regression framework with an individual decision tree, and  $K$  denotes the no of regression trees.

The RF method integrates different tree decisions and supports improving accuracy. To enhance accuracy, the proposed model employs RF and extreme gradient boosting (XG-Boost). For randomization, it employs bagging. Bagging is an abbreviation in Bootstrap aggregation it enhances the method’s reliability and precision.

$$Bagging = \frac{1}{B} \sum_{b=1}^B fb(x') \tag{5}$$

here  $x'$  was the predicted value for the unseen sample, was the total trees,  $b = 1, 2, 3 \dots B$ ; and  $fb =$  Train the DT  $fb$  on  $X_b, Y_b$  [16].

XG-Boost has recently been widely used in numerous research fields, and it has produced favorable outcomes in a variety of machine-learning competitions. XG-Boost differs from GBDT in some aspects (Gradient Boosted Decision Trees). To begin with, the GBDT technique simply uses a first-order Taylor expansion, but the XG-Boost approach increases the loss functions with the Taylor expansion of the second-order. Second, normalization is used by the objective function for preventing overfitting and reducing the technique’s complexity. Third, XG-Boost was very flexible, enabling users to define the objective of optimizations and assessment categories. Nonetheless, the XG-Boost classifier can handle uneven training data effectively by setting class weights and employing AUC as the evaluation parameter.

In summary, XG-Boost is the scalable and adaptable tree structures optimization model capable of managing sparse data, increasing the speed of the algorithm, and reducing processing time and memory for data of large-scale. The XG-Boost algorithm could be formalized as,

Given an n-sample training dataset  $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ,  $x_i \in R^m$ ,  $y_i \in R$ , the objective function could be expressed by,

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{t=1}^T \Omega(f_t) \tag{6}$$

Here,  $l(y_i, \hat{y}_i)$  is the difference between the target  $y_i$  and the prediction  $\hat{y}_i$  and  $f_t$  is the prediction score of the  $t^{\text{th}}$  tree. Based on the Taylor expansions of the objective functions, the estimated loss function may be computed:

$$L^{(t)} \simeq \sum_{i=1}^k \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{7}$$

Here,  $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$  represents every sample’s initial derivative and  $g_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$  signifies every sample’s second derivative, and the loss function only requires the initial and second derivatives of every element of information [14].

To generate model predictions XG-Boost is the most used approach. The main disadvantage of the RF was its slowness. So, to eliminate the problem of speed, XG-Boost was integrated with it to achieve the best outcomes for the proposed research. XG-Boost was a technique utilized in classification approaches to enhance prediction results and minimize errors by lowering bias.

## 7 Experimental analysis

The suggested framework was tested with MATLAB Simulink 2019a. The studies were done out using a PC equipped with an Intel Core i7–10,700 CPU running at 2.9 and 4.8 GHz, 8 GB of RAM, and a 64-bit version of Windows 10. Initially, the dataset was preprocessed and the features were selected using the ALO. The information set is divided between 75% training and 25% testing.

## 8 Dataset description

This information was obtained in real time from different regions around the Madurai district between 2019 and 2020. The dataset varies between January to December. 21,574 samples were collected, which includes 10,879 female samples and 10,695 male samples. The ages of the people range from one to 89. District, geographical area, Month, year, age, gender, illness, death, nausea, bleeds, dividing fever, headache, high temperature fever, mild headaches, bloodstream vomiting, eye pain, nose bleeding, gum bleeding, discomfort in the abdomen, rashes on the skin, muscle pain, joint pain, pain in the body, intense bleeding, dehydration, which organ damage, respiration problem, swelling glands, and tiredness are all variables to consider are among the 29 characteristics in the information set. Out of these 29 features, district, region, and death features are ignored and the remaining 26 features were selected for analysis. The dataset was extracted in the following table (Tables 1, 2, 3 and 4).

Based on the number of cases infected each month, most of the cases are infected from Malaria infection. This malaria infection covers 52% of the infection from the total of 21,574 cases. Then dengue infection covers 23% of the infection the overall cases. Both these malaria and dengue infections were the most infected disease each month. Next to malaria and dengue, pneumonia infection covers 16% of infection, typhoid with 5%, kala-azar with 2%, and Japanese encephalitis and measles each covers 1% separately.

### 8.1 Performance metrics

The performance metrics are the evaluation metrics used in this research for performance analysis.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \% \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

**Table 1** The seasonal influences on illness [11]

Disease/Infection	Type of Disease/Infection	Seasonal Drivers	Description
Pneumonia	Acute	Seasonal non-climatic abiotic environments (e.g., the salinity of water), Seasonal climates (e.g., precipitation, temperature, etc.), and Seasonal co-infections.	Peaks in midwinter it was linked with influenza.
Dengue	Acute	Vector seasonality	Rainy seasons
Malaria	Acute	Vector seasonality	A wide range of seasonality exists, with some locations experiencing high seasonality and others seeing none at all.
Japanese encephalitis	Acute	Vector seasonality	Outbreaks are seasonal in the northern tropical zone; they occur near the end of the rainy season; however, there is no regular pattern in tropical areas.
Measles	Acute	Seasonal exposures and/or Behaviour and/or Contact rates	Seasonality in advancing nations is tied to agricultural cycles and contributes to increased transmission.
Typhoid	Acute	Monitored seasonal incidences with no hypothesis regarding the driver. Seasonal climate (e.g., temperature, precipitation, etc.).	Peaks around July
Kala Azar (Black Fever)	Chronic	Vector seasonality	Sand flies carry the disease; domestic dogs were the primary reservoirs and were exposed to the disease during a certain period.

**Table 2** Dataset details about the number of cases

Month	Infection/Disease	Number of cases
January	Dengue, Typhoid, Malaria, Pneumonia, Kala Azar, Measles	850
February	Japanese Encephalitis, Dengue, Typhoid, Malaria, Pneumonia, Kala Azar, Measles	1546
March	Typhoid, Pneumonia, Malaria, Dengue, Japanese Encephalitis, Kala Azar, Measles	712
April	Dengue, Typhoid, Malaria, Pneumonia	2338
May	Typhoid, Pneumonia, Malaria, Dengue	2691
June	Pneumonia, Malaria, Dengue	2690
July	Typhoid, Pneumonia, Malaria, Dengue	3038
August	Typhoid, Pneumonia, Malaria	3804
September	Pneumonia, Malaria	1214
October	Malaria, Dengue	643
November	Malaria, Dengue	105
December	Malaria, Dengue	1943
Total		21,574

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (11)$$

$$\text{F - Score} = \frac{2TP}{2TP + FP + FN} \quad (12)$$

- TP: It is the number of correct classifications of normal classes is referred to as the real positive value.
- FP: It is a misleading positive value resulting from an inaccurate classification in typical classes.
- TN: It was the total true categorization in anomalous classes that was real false case.

**Table 3** Infection rate with number of cases

Infection/Disease	Number of cases total with %
Malaria	11153 (52%)
Dengue	4892 (23%)
Pneumonia	3463 (16%)
Typhoid	1009 (5%)
Kala Azar	335 (2%)
Measles	302 (1%)
Japanese Encephalitis	268 (1%)
Normal Fever and Cold	152 (below 1%)

**Table 4** Performance analysis comparison of proposed model

Model	Accuracy	Precision	Recall	Specificity	F-Score
ACO-ANN	90.22	89.07	89.81	88.86	90.85
PSO-RF	91.16	89.93	90.43	89.97	92.92
WO-RF	92.49	90.08	92.15	90.12	93.01
ANOVA-SVM	94.95	92.19	93.32	92.30	95.14
Proposed Model	96.17	93.95	95.86	93.23	96.22

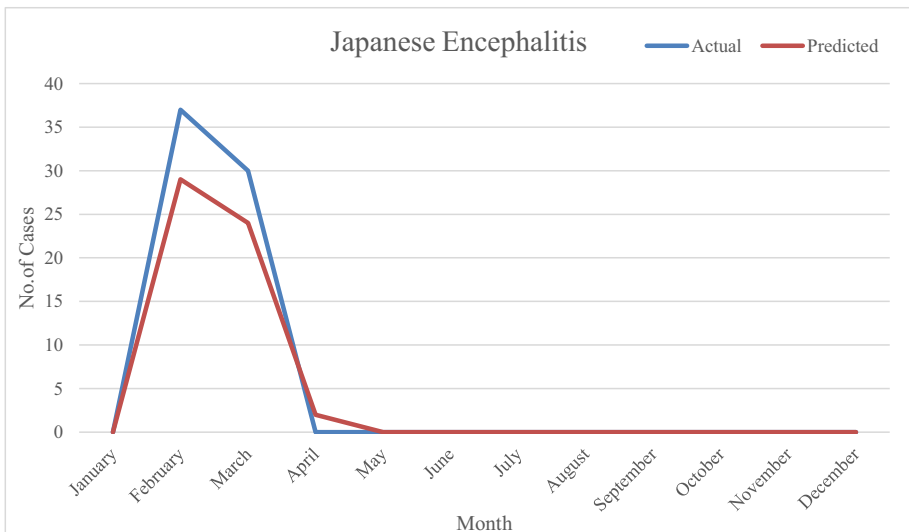
- FN: It was an improper classification in anomalous classes that resulted in the incorrect negative value.

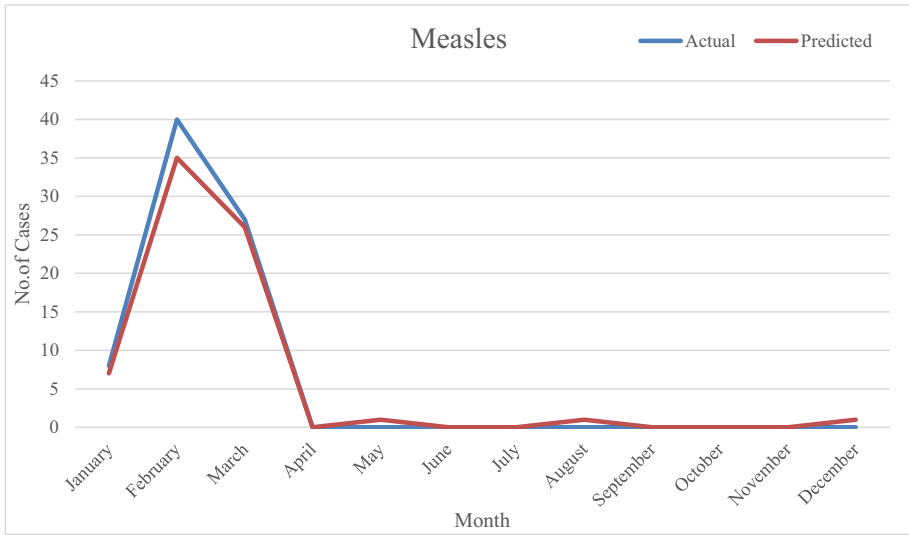
The proposed design was implemented for prediction of a various set of diseases from the given dataset. Based on the training of the model, the model was evaluated with the test data. A disease like Japanese encephalitis was predicted by the proposed model. The difference between the actual and predicted data was compared as shown in Fig. 3. As shown in the figure, the suggested design has close performance related to the actual data.

The Measles disease was predicted by using the evaluation dataset, in the suggested approach. The difference between the actual and predicted data was compared as shown in Fig. 4. As shown in the figure, the suggested design has close performance related to the actual data.

As shown in Figs. 5 and 6, the diseases like Kala Azar and Typhoid were predicted by using the evaluation dataset, in the suggested approach. The difference between the actual and predicted data was compared. As shown in the figure, the proposed model has close performance related to the actual data in both these predictions.

As shown in Figs. 7 and 8, the diseases like Pneumonia and Dengue were predicted by using the evaluation dataset, in the suggested approach. The difference between the

**Fig. 3** Proposed model's predicted results for Japanese encephalitis

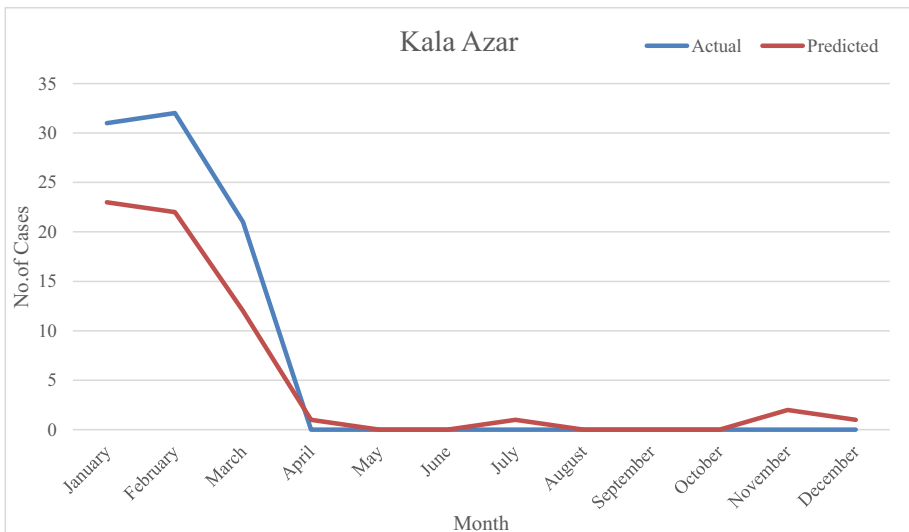


**Fig. 4** Proposed model's predicted results for measles

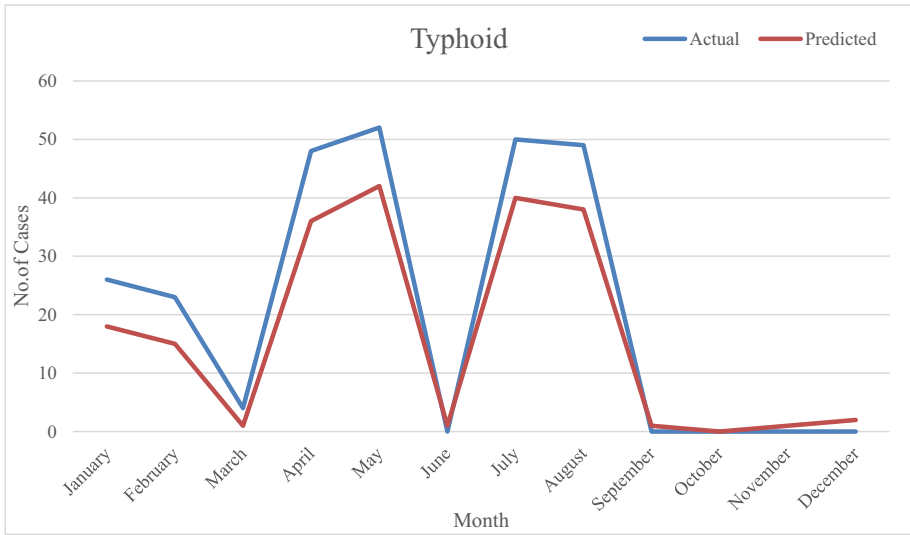
actual and predicted data was compared. As shown in the figure, the proposed model has close performance related to the actual data in both these predictions.

In this research, Malaria is the most infected disease that occurred throughout the year annually. The proposed model has predicted the malaria disease from the test dataset with close performance to the actual data as shown in Fig. 9.

The suggested model was evaluated with the performance metrics like the correctness, exactness, completeness, selectivity, and harmonic mean of precision and recall

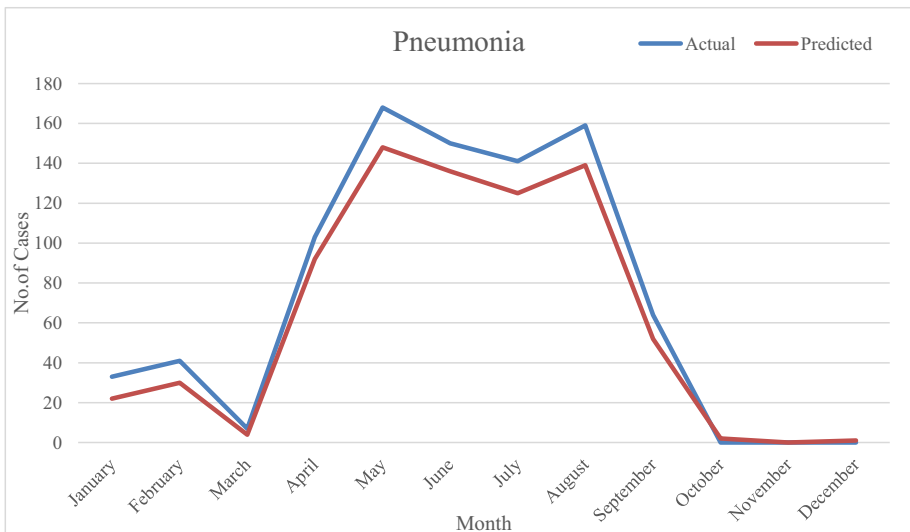


**Fig. 5** Proposed model's predicted results for Kala Azar



**Fig. 6** Proposed model’s predicted results for typhoid

are being referred. As per the experimental analysis, the suggested method has achieved 96.17% accuracy, which is 1.2% to 5.9% improved compared to ACO-ANN, PSO-RF, WO-RF, and ANOVA-SVM models. The precision score of the proposed design was 93.95%, which is 1.7% to 4.8% better than other designs. The recall score of the proposed model was 95.86%, which is 2.5% to 6% enhanced than the other compared models. The specificity of the proposed model was 93.23%, which is 0.9% to 4.3% better than other compared models. The f-score of the proposed design was 96.22%, which is 1.08%



**Fig. 7** Proposed model’s predicted results for pneumonia

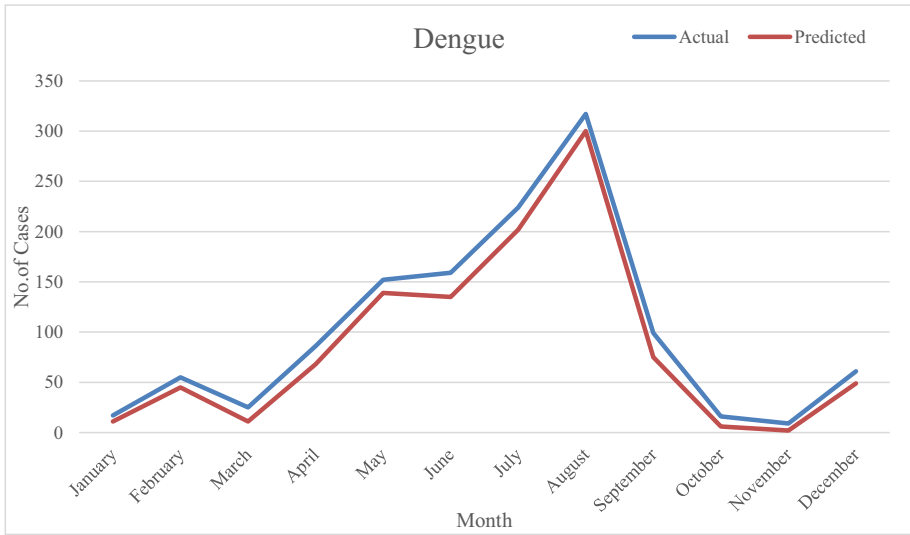


Fig. 8 Proposed model’s predicted results for dengue

to 5.3% improved than other designs. Based on the comparison, the proposed method has better results all the other compared models in terms of all the parameters. The ANOVA-SVM has a close performance related to the proposed model and the least efficiency was achieved by the ACO-ANN model (Fig. 10).

The sensitivity results of the proposed method in Fig. 11 showed in terms of high correctness, exactness, completeness, selectivity, and harmonic mean of precision and recall are being referred and to outperforming other compared models. The Antlion Optimization

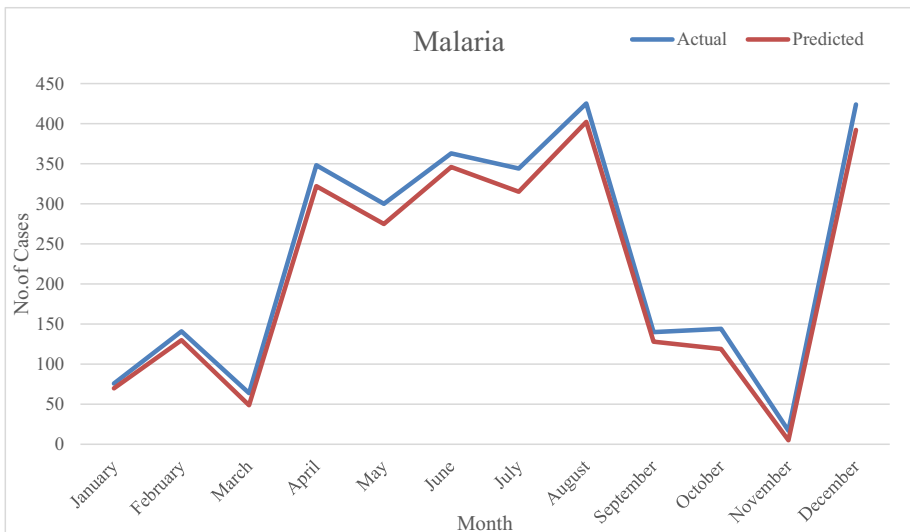
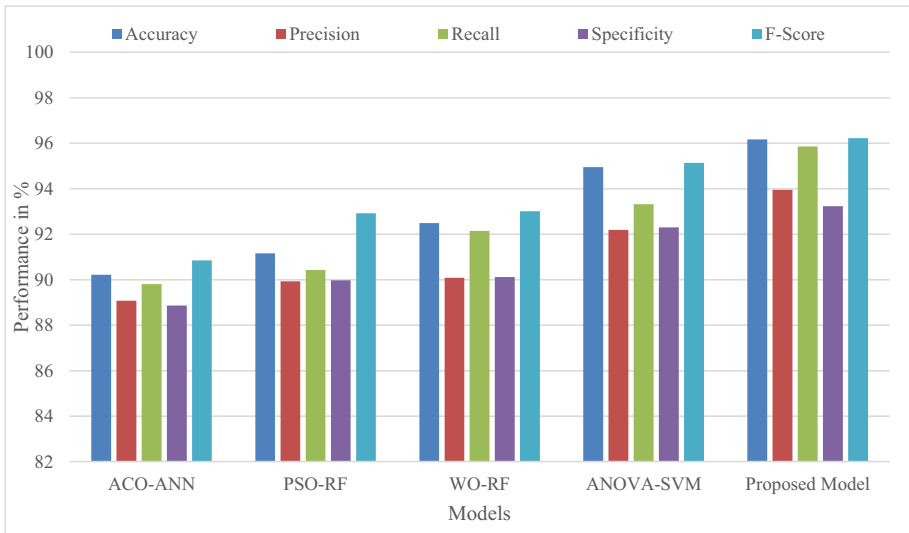


Fig. 9 Proposed model’s predicted results for malaria



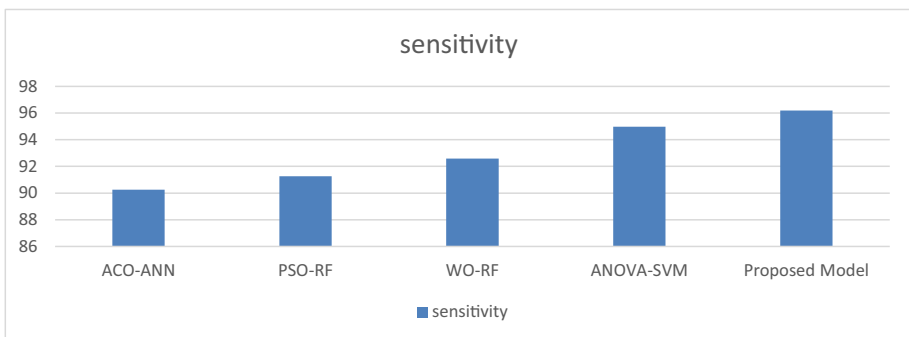


**Fig. 10** Graphical plot for the comparison of performance analysis

Algorithm for feature selection and Random Forest integrated with XG-Boost for classification contributed to the toughness and efficacy of the model. The proposed method demonstrated promising sensitivity results, highlighting its potential for accurate seasonality disease prediction.

### 9 Conclusion

In conclusion, the proposed hybrid machine learning model for predicting seasonal diseases based on real-time data from different regions around the Madurai district has shown promising results. The use of the Antlion Algorithm and Random Forest integrated with XG-Boost for classification has resulted in high performance analysis. Compared to other models such as ACO-ANN, PSO-RF, WO-RF, and ANOVA-SVM, the proposed model has demonstrated superior performance across all parameters. The



**Fig. 11** Graphical plot for the comparison of sensitivity analysis

industrial benefits of this research are significant. By accurately predicting seasonality diseases, such as dengue, malaria, pneumonia, typhoid, and others, based on real-time data, it can enable proactive measures to be taken to prevent or control outbreaks on time. This can lead to improved public health management, reduced healthcare costs, and more efficient allocation of resources. Additionally, the suggested model can be improved further by integrating DL-based techniques to potentially achieve even higher accuracy and efficiency in disease prediction. Furthermore, the proposed model has the potential for broader applications beyond the Madurai district, as it can be adapted to other regions and extended to incorporate climatic and weather changes for disease prediction. This could contribute to early warning systems for disease outbreaks, aiding in proactive planning and preparedness measures. Overall, the proposed model has significant industrial benefits and holds promise for improving the efficiency of disease prediction and management in real-world scenarios.

**Acknowledgements** There is no acknowledgement involved in this work.

**Authorship contributions** There is no authorship contribution.

**Funding** No funding is involved in this work.

## Declarations

**Ethics approval and consent to participate** No participation of humans takes place in this implementation process.

**Human and animal rights** No violation of Human and Animal Rights is involved.

**Conflict of interest** Conflict of Interest is not applicable in this work.

## References

1. Arqum M, Singh A, Cherifi H (2020) Impact of seasonal condition on vector-borne epidemiological dynamics. *IEEE Access* 8:94510–94525
2. Bhatnagar S, Lal V, Gupta SD, Gupta OP (2012) Forecasting incidences of dengue in Rajasthan, using time series analyses. *Indian J Public Health* 56(4):281
3. Davi C et al (2019) Severe dengue prognosis using human genome data and machine learning. *IEEE Trans on Biomed Eng* 66(10):2861–2868
4. Dutta P, Paul S, Obaid AJ, Pal S, Mukhopadhyay K (2021) Feature selections based artificial intelligence technique for the predictions of COVID like diseases. *J Phys Conf Ser* 1963(1):012167
5. Gambhir S, Malik SK, Kumar Y (2017) PSO-ANN based diagnostics model for the early detections of dengue diseases. *New Horizons Transl Med* 4(1–4):1–8
6. Gambhir S, Malik SK, Kumar Y (2018) The diagnosis of dengue disease: an evaluation of three machine learning approaches. *International Journal of Healthcare Information Systems and Informatics (IJHISI)* 13(3):1–19
7. Gothai E, Natesan P, Rajalaxmi RR, Vignesh T, Srinithy K, Balaji TV (2021) Predictive analysis in determining the dissemination of infectious disease and its severity. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1556–1562). IEEE
8. Grampurohit S, Sagarnal C (2020) Disease predictions using machine learning algorithm. In *International Conferences for Emerging Technology (INCET)*, pp. 1–7
9. Indhumathi K, Kumar KS (2021) A review on prediction of seasonal disease based on climate change using big data. *Mater Today: Proceed* 37:2648–2652
10. Iqbal N, Islam M (2017) Machine learning for dengue outbreak predictions: an outlook. *Int J Adv Res Comput Sci* 8(1):93–102

11. Martinez ME (2018) The calendar of epidemics: seasonal cycle of infectious diseases. *PLoS Pathog* 14(11):e1007327
12. Miyashita K, Nakatani E, Hozumi H, Sato Y, Miyachi Y, Suda T (2021) Risk factors for pneumonia and death in adult patients with seasonal influenza and establishment of prediction scores: a population-based study. In *Open Forum Infectious Diseases* 8(3):ofab068. Oxford University Press, New York
13. Mussumeci E, Coelho FC (2020) Large-scale multivariate forecasting model for dengue-LSTM versus random forest regressions. *Spat Spatio-Temporal Epidemiol* 35:100372
14. Nasiri H, Alavi SA (2021) A novel framework based on deep learning and ANOVA feature selections method for diagnosis of COVID-19 cases from chest X-ray Images, arXiv preprint arXiv:2110.06340
15. Rodó X, Pascual M, Doblas-Reyes FJ, Gershunov A, Stone DA, Giorgi F, Hudson PJ, Kinter J, Rodríguez-Arias MÀ, Stenseth NC, Alonso D (2013) Climate change and infectious diseases: can we meet the needs for better prediction? *Climatic change* 118:625–640
16. Sandhu AK, Batth RS (2021) Software reuse analytics using integrated random forest and gradient boosting machine learning algorithms. *Softw Pract Exp* 51(4):735–747
17. Sharma N, Dev J, Mangla M, Wadhwa VM, Mohanty SN, Kakkar D (2021) A heterogeneous ensemble forecasting model for disease prediction. *New Generation Computing*, pp 1–15
18. Tran TQ, Sakuma J (2019) Seasonal-adjustment Based Feature Selections Method for Predicting Epidemic with Large-scale Search Engine Log. In *Proceeding of the 25th ACM SIGKDD International Conferences on Knowledge Discovery & Data Mining*, pp. 2857–2866
19. Zawbaa HM, Emary E, Parv B (2015) Feature selection based on antlion optimization algorithm. In *2015 Third world conference on complex systems (WCCS)* (pp. 1–7). IEEE

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.