Bulletin of the National Research Centre

RESEARCH **Open Access**

Prediction of measles patients using machine learning classifiers: a comparative study

Robert Gyebi[1*], Gabriel Asare Okyere[2], Emmanuel Kwaku Nakua[1], Franklin Aseidu-Bekoe[3], Jane Serwaa Akoto Nti[4], Emmanuel Owusu Ansah[4] and Felix Agyemang Opoku[5]

Abstract

Background Measles has high primary reproductive number, extremely infectious and ranked second to malaria in terms of disease burden in Ghana. Owing to the disease's high infectious rate, making early diagnosis based on an accurate system can help limit the spread of the disease. Studies have been conducted to derive models to serve as preliminary tools for early detection. However, these derived models are based on traditional methods, which may be limited in terms of model sensitivity and prediction power. This study focuses on comparing the performance of five machine learning classification techniques with a traditional method for predicting measles patients in Ghana. The study was an analytical cross-sectional design of suspected measles cases in Ghana.

Results The performance of six classifiers were compared and the random forest (RF) model demonstrated better performance among other models. The RF model achieved the highest sensitivity (0.88) specificity (0.96), ROC (0.92) and total accuracy (0.92).

Conclusions Our findings showed that, despite all the six methods had good performance in classifying measles patients, the RF model outperformed all the other classifiers in terms of different criteria in prediction accuracy. Accordingly, this approach is an effective classifier for predicting measles in the early stage.

Keywords Measles, Prediction, Machine learning, Classification techniques, Prediction accuracy

Background

Measles is a respiratory disease caused by a paramyxovirus, which is highly contagious CDC: An introduction to measles (2019). The virus affects the respiratory mucosa and is transmitted by droplets released when an infected individual coughs or sneezes. People who have no immunity against the virus (thus, between seventy-five and ninety-five percent) become infected when they come into contact with the virus. Almost all persons who are infected become clinically ill (Misin et al. 2020).

The worldwide measles case count by the year 2015 had surpassed 9.7 million, with 254,928 cases recorded throughout the World Health Organization's (WHO) six regions (Kuehn 2021). An estimated 134,200–140,000 measles deaths were recorded in 2018, and the most affected were children under 5 years of age (Kuehn 2021). As a result of effective vaccination programs, the figure was 73% lower than it was in 2000. In the WHO regions in Africa, a case-based surveillance conducted from the year 2013 to 2016 revealed that a total of 176,785 measles

*Correspondence:
Robert Gyebi
rgyebi1@st.knust.edu.gh
[1] Department of Epidemiology and Biostatistics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[2] Department of Statistics and Actuarial Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[3] Public Health Division, Ghana Health Service, Accra, Ghana
[4] School of Medicine and Dentistry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[5] Department of Occupational and Environmental Health and Safety, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

cases were confirmed (Stephenson 2021). In Ghana, measles is a significant contributor to child mortality. According to The Ghana Health Assessment Project, measles is second to malaria in terms of disease burden. It accounted for about 8.8% and 9.3% of persons admitted to Ghana's health facilities (Kissi et al. 2022).

Identifying the most relevant demographic factors related to measles is essential since it substantially contributes to infant mortality. These factors include gender, age group, region, settlement type and vaccination status. By increasing the fraction of measles patients detected at an early stage, early identification of the disease can play a crucial role in improving patients survival (Rao et al. 2022). Traditional classification approaches, such as logistic regression has been routinely utilized to detect measles cases in various medical respects. While these models might provide straightforward interpretations, they often fail to account for complicated relationships between variables. As a result, newly developed models with the lowest prediction error, precise and reliable approach for early patient diagnosis are required. Most contemporary medical diagnostic methods are built on classification, and numerous researchers have altered them to enhance precision.

Machine learning techniques have recently gained popularity and are now widely utilized in various fields including medicine, particularly in classification issues (Saladi et al. 2023; Hasan et al. 2022). These techniques enhance their performance with time and help clinicians identify new patients more accurately by enhancing sensitivity and decision-making (Hasan et al. 2022). Even though the primary goal of these models is to discover influential factors and their relationships, they may also be used to forecast and quantify effects (Sharma and Raghava 2022; Kumar et al. 2023). In most researches (Mirzaei and Adeli 2022; Allugunti 2022), various machine learning algorithms have been introduced to predict diverse outcomes. Machine learning techniques such as Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Nearest Neighbor, AdaBoost, Support Vector Machine (SVM) and Multilayer Perceptron are used (Gu and Lu 2021; Charbuty and Abdulazeez 2021; Car et al. 2020). Although numerous studies have demonstrated that data mining approaches perform better than traditional techniques in terms of accuracy and error rates, this excellence does not occur in all data sets (Yang et al. 2020), and there are discrepancies among different researches. As a result, investigating and comparing the performance of various methods in diverse data sets is essential.

This study aims to compare five machine learning techniques: Naive Bayes, Support Vector Machines, Artificial Neural Networks, Decision Trees, and Random Forests, as well as a traditional method (Logistic Regression) in predicting measles in Ghana to separate people with measles from healthy people.

Literature review

Machine learning algorithms are becoming increasingly popular in various fields due to their ability to analyze large amounts of data and identify patterns that can aid in decision making. In this literature review, we will discuss some of the popular machine learning algorithms, their applications in infectious disease outbreak prediction, and other related fields.

Naive Bayes

Naive Bayes is a popular and efficient machine learning algorithm used for classification tasks. It calculates the probability of each class based on the probability of each feature occurring in that class and selects the class with the highest probability as the prediction. Naive Bayes is particularly useful for text classification, spam filtering, and infectious disease outbreak prediction. In a recent study conducted in Mexico, Naive Bayes showed the highest specificity of 94.30% in predicting COVID-19 infections (Muhammad et al. 2021).

Random forest

Random forest is an ensemble learning method that uses multiple decision trees to make predictions. It constructs a multitude of decision trees at training time and outputs the mode or mean prediction of the individual trees. Random forest has been applied to infectious disease outbreak prediction in several studies. In a study conducted to predict COVID-19 using machine learning algorithms, random forest had the best precision of 94.99% (Yadav 2021).

Decision trees

Decision Trees are a simple yet powerful machine learning algorithm that can be used for infectious disease outbreak prediction. They work by recursively partitioning the feature space into smaller regions based on the values of the input features. Once the tree is constructed, it can be used to make predictions by traversing the tree from the root to a leaf node that corresponds to a particular prediction. Decision Trees have achieved the best accuracy of 99.93% in recent studies conducted in the analysis of medical diseases (Jijo 2021).

Support vector machines (SVMs)

SVMs are a powerful and widely used machine learning algorithm that can be used for both classification and regression tasks. They work by finding a hyperplane that separates the data into different classes in a way that maximizes the margin between the two classes. SVMs can handle both linearly separable and nonlinearly separable data using different kernel functions that transform the data into a higher-dimensional space. In precision psychiatry, SVMs have been used for applications that involve diagnosis and prognosis prediction of brain diseases like Alzheimer's disease, schizophrenia, and depression Pisner and Schnyer (2020).

Artificial neural networks (ANNs)

ANNs are a family of machine learning algorithms inspired by the structure and function of biological neurons in the brain. They consist of a large number of interconnected processing nodes that are organized into layers. ANNs can be trained using a variety of learning algorithms, such as backpropagation, to adjust the weights of the connections between neurons in order to minimize a loss function. ANNs have been applied to various fields, including the detection and segmentation of contrast-enhancing tumors in neuro-oncology Kickingereder et al. (2019).

Generalized linear models (GLMs)

GLMs are a flexible and widely used class of statistical models that extend the linear regression framework to handle a wider range of response variables. GLMs use a link function to relate the linear predictor to the response variable, which can have a non-normal distribution. GLMs can be fit using maximum likelihood estimation, which involves finding the parameter values that maximize the likelihood of the observed data given the model. GLMs have been used for predicting gene expression patterns in China Liu et al. (2019).

Gini index

The Gini index is a commonly used measure of the inequality in a distribution, and it can be used to evaluate the diagnostic share of each predicting variable in a predictive model. The Gini index measures the degree of dissimilarity or diversity among a set of values, and in the context of predictive modeling, it can be used to assess the importance of each variable in the model. When comparing the Gini Index values between different features, a higher Gini Index value indicates a better split. This means that the feature with the highest Gini Index value is more effective at separating the data into groups with different target values. Therefore, features with higher Gini Index values are considered more important in the decision tree algorithm.

Methods

Dataset description

We used a secondary data set that contained six attributes such as gender, age group, region, settlement, vaccination status, and measles status from Ghana health service for the study. The data was recorded in a case-based form after a clinical diagnosis of suspected individuals who report at the health facility. Furthermore, the clinical and laboratory information of the suspected individuals were available and accessible from their medical records. Regardless of age and gender, individuals who were clinically diagnosed were selected. In this study, the suspected individual was defined as a measles patient who presented signs and symptoms of measles at the health facility and was clinically diagnosed by a physician and was yet to be confirmed by the laboratory outcome after collecting specimens for testing. The patients with the following criteria were either included or excluded in the study:

Inclusion criteria

(i) All suspected measles cases reported at the health facility.
(ii) Cases with the adequate sample. In medical laboratory specimen sample collection, an adequate sample size refers to the amount of biological material (e.g., blood, urine, tissue) collected from a patient that is sufficient for the intended laboratory testing or analysis. In addition to the amount of biological material collected, an adequate sample size may also depend on other factors, such as the quality and condition of the sample, the storage, and transportation methods used.
(iii) Cases with confirmed laboratory outcome.

Exclusion criteria

(i) Cases with inadequate samples for laboratory testing
(ii) Not yet confirmed laboratory results.
(iii) Patients with missing information were excluded from the study.

The demographic characteristics of the study participants were age, gender, region and type of settlement. Other categorical variables included in the study was vaccination status (vaccinated, not vaccinated and unknown vaccination status).

Methodologies

In this study, we aimed to predict measles patients using several machine learning classifiers, including Generalized Linear model, Random Forest (RF), Decision Tree

Table 1 ANN parameters used for classification

Parameter	Value
Training	80%
Testing	10%
Validation	10%
Hidden layer	5
Input neurons	5
Output neurons	1
Algorithm	backpropagation
Activation function	Sigmoid
Repetitions	2
Lifesign	minimal
stepmax	1000
Learningrate	0.01
learningrate.limit	(0.1, 0.9)

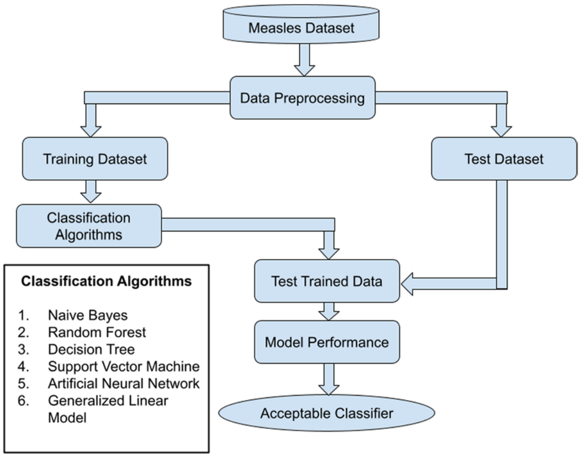

Fig. 1 A flow chart of the methodology showing how the data was processed using various classification algorithms

(DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes. We compared the performance of these classifiers to determine their effectiveness for this task. We utilized pre-existing built-in packages for these algorithms in the Python programming language in the Jupyter notebook environment and ”Rstudio”. We also included a table outlining the specific ANN parameters used for classification (Table 1) and a figure illustrating the techniques used (Fig. 1).

(i) The sigmoid function in the ”neuralnet” package in ”R”, has a default **alpha** value of 1. Since the ”neuralnet” package in ”R” does not provide an **alpha** parameter, the **alpha** value is set as 1 by default.

(ii) *Repetition:* This refers to the number of times the training process is repeated. It represents the number of iterations or epochs during which the neural network’s weights are updated based on the training data.

(iii) *Lifesign:* Lifesign parameter in the “neuralnet” package in “R” controls the frequency of progress updates or messages during the training process. When set to ”minimal,” it means that only essential progress information will be displayed, such as the current step or iteration.

(iv) *Stepmax:* This is a parameter in the “neuralnet” package in “R” that specifies the maximum number of steps or iterations allowed during the training process. It determines the maximum number of times the weights of the neural network will be adjusted based on the training data.

(v) *Learning rate:* The learning rate parameter determines the step size or magnitude of weight updates during the training process. In the given parameter setting, the learning rate is set to a constant value of 0.01 throughout the training.

The following steps were used to classify the measles patients:

1 Data pre-processing was the process of converting raw data into clean data that can be analyze. *Handling imbalance class* The current measles dataset contains 1,696 patients in the negative class, while the positive class is represented by only 78 patients. However, this data may be limited or biased, potentially compromising the accuracy and generalizability of machine learning models trained on it. To address these issues, we propose generating synthetic data to overcome the limitations of the available data and improve model performance (Zhu et al. 2017; Wang and Yao 2012; Malhotra and Kamal 2019). Class imbalance, a situation where one class has significantly fewer observations than the other, can lead to models being biased towards the majority class during training. In machine learning tasks with a small number of positive cases, generating synthetic data can help alleviate this problem. We employed the random oversampling method to generate synthetic data for the minority class and balance the dataset before training. This method involves randomly creating instances from the minority class and adding them to the training dataset. We generated a total of 2536 new observations, with 1268 being classified as positive cases and the remaining 1268 as negative cases. Our results demonstrate that this approach effectively balances the data and leads to more accurate and generalizable machine learning models, even

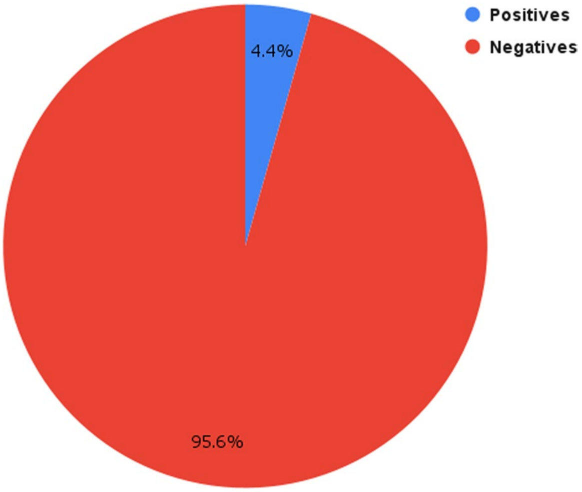

Fig. 2 Unbalanced data (Original dataset): Total = 1,797; Positives = 78; Negatives = 1,696

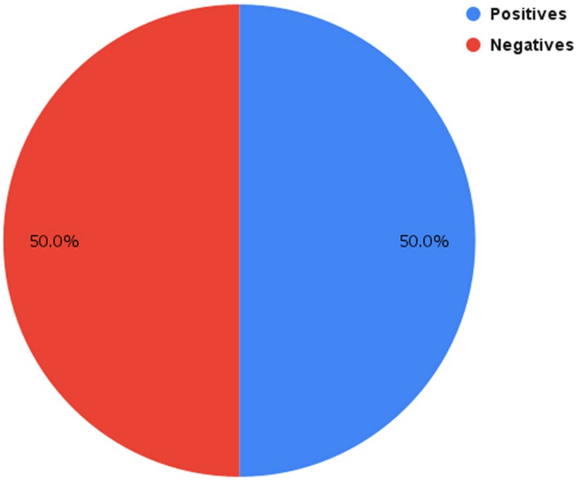

Fig. 3 Balanced data: Total = 2536; Positives = 1,268; Negatives = 1,268

with limited or biased datasets. Figures 2 and 3 are illustrations of the data.

2 The tidy data was subjected to a percentage split.

3 The tidy dataset was divided into training (75%, n = 1,902) and test (25%, n = 634) datasets by the percentage split.

Results

Data description

In this study, a total of 1,797 suspected cases of measles were obtained from the measles database. The study population consisted of 1,016 males (56.5%) and 781 females (43.5%) with ages ranging from 1 month to 29 years and above, with a mean age of 54.29 months. Table 2 displays the relationship between measles status and the demographic characteristics of the study participants. A chi-squared test was used to examine the relationship between the categorical variables. Although males made up the majority of the study population, gender was not significantly associated with measles status (p = 0.708). The majority of the study participants were aged 1–4 years (56.7%, p<0.001). Measles was most suspected among individuals from the Upper West region of Ghana (12.9%), followed by individuals from the Eastern region of Ghana (11.4%). Conversely, cases of measles were least suspected among individuals from the North East region of Ghana (0.8%). Notably, the majority of the study population was from rural settlements (69.8%) and region of patients and settlement type were statistically significant (p<0.001). Additionally, 67.7% of the study patients had been vaccinated against measles, while 12.3% were unvaccinated, and 20% could not recall their vaccination status. These findings suggest that region of residence and settlement type may be associated with measles susceptibility, and vaccination status remains an important factor in the prevention and control of measles outbreaks in Ghana.

Models performance

Figure 4 illustrates a comparison of the confusion matrix of all the classification techniques, including the traditional Generalized Linear Model (GLM) method used. According to the results in Fig. 4, Random Forest (RF) achieved a True Positive Rate (TPR/ Sensitivity) of 0.883, which was higher than Decision Tree (DT), which achieved 0.787 and the traditional GLM method, which also achieved a TPR of 0.738. The RF model was highest among the other models used in this study in terms of sensitivity. Again, the random forest method achieved a specificity (FPR) of 0.964, far higher than the specificity (FPR) of the other classification techniques. The random forest technique achieved the highest precision (PPV) compared with the other classifiers. The random forest model achieved the highest ability to recall (NPV) compared with the recall (NPV) of the other models. The Area Under Curve (AUC) of the ROC is the prediction power of the model, which the RF model had the highest predictive power (0.923) compared with the other classification techniques (Figs. 4 and 5).

According to the results in Fig. 6, the RF model had the highest total accuracy (92.11%), followed by DT (78.39%) and then the traditional GLM method (72.87%). Based on the total accuracy, the RF method predicted better than the other models.

Table 2 Demographic Characteristics of patients

Demographic characteristics	Frequency	Proportion (%)	X-Squared	p-value
Gender			0.395	0.5299
Male	1016	56.5		
Female	781	43.5		
Age group			90.179	0.001
<9 months	167	9.3		
9 - 11 months	136	7.6		
1 - 4 years	1019	56.7		
5 - 9 years	296	16.5		
10 - 14 years	86	4.8		
15 - 19 years	29	1.6		
20 - 24 years	19	1.1		
25+ years	44	2.4		
Missing	1	0.1		
Region of residence			26.165	0.036
Upper West	231	12.9		
Eastern	205	11.4		
Bono	183	10.2		
Volta	170	9.5		
Greater Accra	168	9.3		
Upper East	151	8.4		
Ashanti	146	8.1		
Bono East	115	6.4		
Western	91	5.1		
Central	87	4.8		
Northern	71	4.0		
Ahafo	52	2.9		
Oti	41	2.3		
Western North	36	2.0		
Savannah	36	2.0		
North East	14	0.8		
Type of settlement			7.897	0.005
Urban	542	30.2		
Rural	1,255	69.8		
Vaccination status			17.629	0.001
Vaccinated	1,216	67.7		
Not vaccinated	221	12.3		
Unknown	360	20.0		

Figure 7 compares the AUCs (Area Under the ROC Curve) of different classifiers. To plot the ROC curve, we first calculate the True Positive Rate (TPR) and False Positive Rate (FPR) at various classification thresholds. The FPR is defined as the ratio of negative samples that are incorrectly classified as positive to the total number of negative samples. We vary the classification threshold from 0 to 1 and calculate the TPR and FPR for each threshold by comparing the model's predictions to the true labels of the test set.

The AUC of the Random Forest (RF) model was found to be higher than the AUC of the other models. This indicates that the RF model had higher predictive power compared to the other models. Specifically, the AUC of the RF model was 92.3%, while the AUC of the traditional Generalized Linear Model (GLM) method was 72.8%.

Evaluation of the RF model

Table 3 presents the first ten observations of measles cases used to evaluate the performance of the predicted

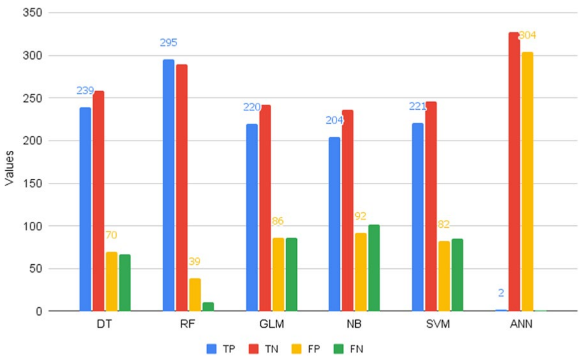

Fig. 4 Confusion matrix for measles dataset

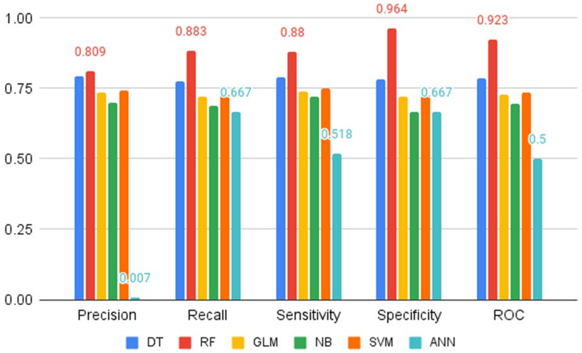

Fig. 5 Performance parameters for measles dataset

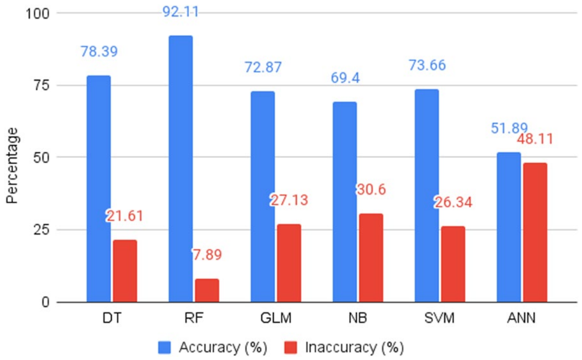

Fig. 6 Accuracy and inaccuracy for measles dataset

model. For the first patient, the laboratory outcome indicated a positive measles status (actual = 1), but the model's predicted probability of testing positive was only 0.002 (0.2%). This indicates an error in the model's prediction, resulting in a negative predicted outcome (converted predicted probability status = 0) even though the first patient was an actual positive case. However, in the subsequent nine observations, the model accurately predicted the measles status when compared to the actual laboratory results(Table 3).

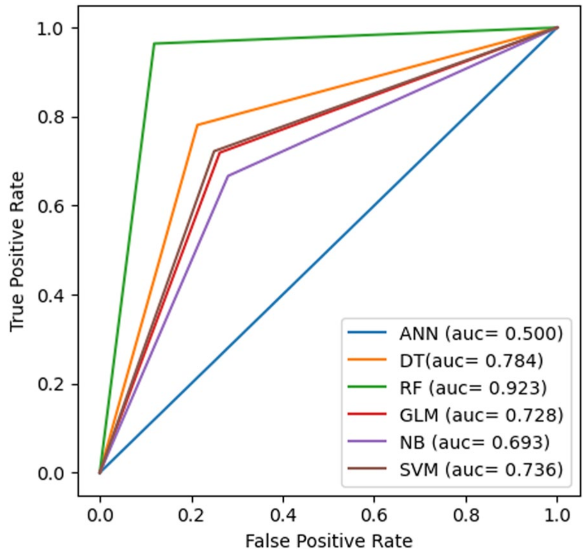

Fig. 7 Area Under the curve of the classifiers

Table 3 Evaluation of the RF Model

Case	Actual Measles Status	Model Predicted Probability	Converted predicted probability status
1	1	0.002	0
2	1	0.755	1
3	1	0.755	1
4	1	0.755	1
5	1	0.755	1
6	1	0.755	1
7	1	0.755	1
8	1	0.755	1
9	1	0.755	1
10	1	0.935	1

Confusion matrix of the RF model

The True Positive (295) and True Negative (289) of the confusion matrix was high relative to the False Positive (39) and False Negatives (11) of the predicted model (Table 4).

Performance of the RF model

The overall performance or accuracy of the predicted model was estimated to be 92.11%. The model correctly classified about 88.11% as positive (sensitivity), while about 96.41% was the proportion of individuals correctly classified as negative (specificity). The model's probability

Table 4 Confusion matrix of the RF model

	Predicted Probability	
Prediction	Negative	Positive
Negative	289	11
Positive	39	295

of predicting an individual being infected with measles was 0.9633 (96.33%), whereas the probability of predicting an individual being negatively infected with measles was 0.8832 (88.32%). The overall predictive power of the model (AUC) was estimated to be 92.3% (Table 5).

Variable of importance

The results in Table 6 show the variable importance measures for the decision tree model, including Mean Decrease Accuracy and Mean Decrease Gini. The Province of Residence feature has the highest Gini Index value for both metrics (322.9542 for Mean Decrease Accuracy and 145.56812 for Mean Decrease Gini), indicating it is the most effective feature for splitting the data in the decision tree algorithm. Agegroup (312.9170) and Urbanrural (201.7443) also have relatively high Gini Index values, while Sex (191.4150) and Vaccine Status (141.6461) have lower Gini Index values. These results suggest that Province of Residence, Agegroup, and Urbanrural are more important features in the model than Sex and Vaccine Status. Figure 8 presents a graphical representation of the Mean Decrease Gini for the model.

Discussion

The study aimed at a comprehensive comparison of five machine learning techniques of NB, RF, DT, SVM, ANN and a traditional method (GLM) for the prediction of measles to distinguish people with measles from healthy people in Ghana.

For all five methods, the performance criteria were similar among classifiers; however, they were derived from different algorithm approaches. Based on the total accuracy, only one of the six classifiers tested showed a total accuracy value lower than 0.6 (ANN with total accuracy of 0.518). In other words, in predicting the classes for measles, all the classification methods provided varied accuracy. However, the total accuracy of the RF model was outrageously higher (0.921) than the other techniques. Although RF and NB were the most common algorithms used in practice (El Khoury et al. 2019; Asadollahi et al. 2013; Sanson et al. 2019; Xu et al. 2017), the RF model was used for additional analysis. In the 75%, 25% scenarios, the sensitivity varied from at least 0.518 in ANN to 0.88 in the RF model. In the case of specificity, however, the RF model performed better than other models (0.964), the NB and ANN models performed were poor (0.667). This quality also remains valid for PPV. In other words, RF is the best model based on the NPV and PPV criteria.

Table 6 Variable Importance Measures: Comparison of Metrics

Variables	Mean Decrease Accuracy	Mean Decrease Gini
Sex	191.4150	36.17658
Agegroup	312.9170	86.20044
Province of residence	322.9542	145.56812
Urbanrural	201.7443	39.32388
Vaccine status	141.6461	48.28824

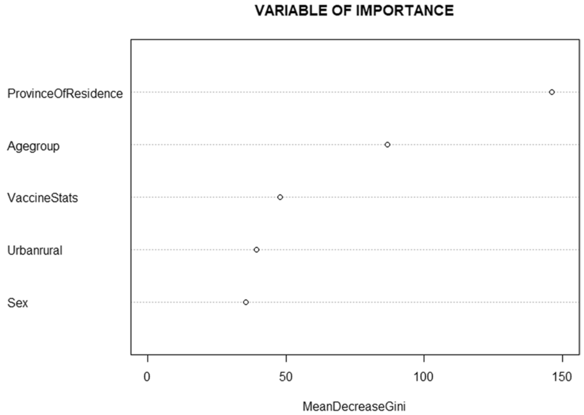

Fig. 8 Mean decrease Gini of the Model

The maximum sensitivity and NPV value belonged to RF. However, the RF model outperformed other models based on the other reliability indices, and it is more effective than NB, DT, SVM, ANN and GLM. Our

Table 5 Performance of the Random Forest model

Data Type	Acc	Sen	Spec	PPV	NPV	AUC
Training data	0.9048	0.8547	0.9541	0.9482	0.8698	-
Test data	0.9211	0.8811	0.9641	0.9633	0.8832	0.9230

Acc- Accuracy; Sen- Sensitivity; Spec- Specificity; PPV- Positive Predictive Value; NPV- Negative predictive Value; AUC- Area Under the Curve finding indicated region as the highest risk factor associated with measles prediction. This result is consistent with the findings (Rao et al. 2021; Portnoy et al. 2019; Ristić et al. 2019). Population size, cultural difference, nutrition, socioeconomic status, public health education and community awareness play significant roles in terms of measles virus infection, which could account for the inter-regional variation observed in the present study.

According to the finding, age group was the second important variable in predicting measles patients. Measles is more likely to occur in the 20 years and above age group (Rao et al. 2021; Portnoy et al. 2019; Ristić et al. 2019). Our analysis indicates that patients in their late twenties were at a high risk of measles. Measles studies conducted elsewhere indicated higher prevalence in children less than 5 years of age (Gujar et al. 2021; Misin et al. 2020; Chen and Whitehead 2021; Sanyaolu et al. 2019; Shen et al. 2020; Benn et al. 2020; Patel et al. 2020). These findings were inconsistent with results from this study which indicated a higher incidence of measles among individuals aged five years and above. The inconsistencies of the findings in this study and studies done elsewhere may be as a result of either the previous studies recruited children under 5 years (Gujar et al. 2021; Misin et al. 2020; Chen and Whitehead 2021; Sanyaolu et al. 2019; Shen et al. 2020; Benn et al. 2020; Patel et al. 2020) (primarily retrospective studies) or were hospital-based investigations.

The third factor that influenced the prediction was vaccination status. This study has revealed that a person's chance of being infected with measles might depend on his or her vaccination status. This finding was consistent with studies done elsewhere which showed that vaccination is intended to reduce infection and mortality rates since it passively increases the adaptive immune system (increase in memory cell production), thereby boosting the immune competency level of an individual (Messina et al. 2019; Geckin et al. 2022; Hayman 2019; Akindele 2022) and this significantly makes unknown vaccination status a potent risk factor for the incidence of measles in Ghana and most African countries where measles used to be rife in the pre-vaccination era.

There were several limitations to our study. The data was missing crucial information about the signs and symptoms of measles. Unfortunately, data on the persons' or caregivers' socioeconomic factors is also limited, and a significant percentage of measles patients with unclear vaccination status remains a key concern that could have influenced and biased the results of this study. Again, incomplete and pending test results of measles were significant issues that could affect the outcome of the present study's analysis.

Conclusions

In this study, we aimed to compare the performance of five machine learning algorithms and a traditional technique for predicting measles patients. Our findings suggest that the random forest (RF) algorithm was the most effective model for predicting measles, based on multiple criteria. Specifically, the RF algorithm demonstrated superior accuracy, precision, recall, sensitivity, and specificity compared to the other models. Our results contribute to the existing literature on the use of machine learning algorithms for disease prediction and demonstrate the potential of these techniques in the context of measles elimination.

Overall, our findings provide valuable insights for policy makers and practitioners in the health sector who are interested in developing predictive models to support measles elimination efforts. Specifically, the use of the RF algorithm could lead to more accurate and effective predictions of measles patients, allowing for more targeted interventions and resource allocation. However, it is important to note that the results of this study are based on a specific dataset and context, and further research is needed to evaluate the generalizability of these findings to other populations and settings.

Abbreviations

RF	Random forest
ROC	Receiver operating characteristic
WHO	World Health Organization
NB	Naive Bayes
DT	Decision tree
SVM	Support vector machine
ANN	Artificial neural network
TPR	True positive rate
FPR	False positive rate
PPV	Positive predictive value
NPR	Negative predictive value
FN	False negative
TP	True positive
GLM	Generalized linear model
AUC	Area under curve

Acknowledgements

The authors are grateful to the disease surveillance department of the Ghana Health Service, Mrs. Elizabeth Alhassan and Mr. Robert Ossom that greatly assisted the research and development.

Author Contributions

RG performed the protocol, analyzed the data, and wrote the original article. JSAN, EOA collected and clean the data. JSAN, EOA and FAO discussed part of the study finding GAO, EKN and FAB did the conception and design of the work, and make critical revisions to the final manuscript. All authors read and approved the final manuscript.

Funding

No funding was obtained for this study.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was carried out following the approval of the local Ethics Committee. All the cases provided their consent to participate in this study.

Consent for publication

Written consent to publish this information was obtained.

Competing interests

The authors declare that they have no competing interests.

Received: 17 February 2023 Accepted: 28 June 2023
Published online: 26 July 2023

References

Akindele NP (2022) Updates in the epidemiology, approaches to vaccine coverage and current outbreaks of measles. Infect Dis Clin 36(1):39–48

Allugunti VR (2022) Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int J Eng Comput Sci 4(1):49–56

Asadollahi S, Fakhri M, Heidari K, Zandieh A, Vafaee R, Mansouri B (2013) Cigarette smoking and associated risk of multiple sclerosis in the Iranian population. J Clin Neurosci 20(12):1747–1750

Benn CS, Martins CL, Andersen A, Fisker AB, Whittle HC, Aaby P (2020) Measles vaccination in presence of measles antibody may enhance child survival. Front Pediatr 8:20

Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V (2020) Modeling the spread of covid-19 infection using a multilayer perceptron. Computational and mathematical methods in medicine, 2020

CDC: An introduction to measles (2019) https://www.cdc.gov/measles/downloads/introtomeaslesslideset.pdf Accessed 23 Nov 2021

Charbuty B, Abdulazeez A (2021) Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends 2(01):20–28

Chen CC, Whitehead A (2021) Emerging and re-emerging infections in children: Covid/mis-c, zika, ebola, measles, varicella, pertussis immunizations. Emerg Med Clin 39(3):453–465

El Khoury Y, Collongues N, De Sèze J, Gulsari V, Patte-Mensah C, Marcou G, Varnek A, Mensah-Nyagan AG, Hellwig P (2019) Serum-based differentiation between multiple sclerosis and amyotrophic lateral sclerosis by random forest classification of ftir spectra. Analyst 144(15):4647–4652

Geckin B, Föhse FK, Domínguez-Andrés J, Netea MG (2022) Trained immunity: implications for vaccination. Curr Opin Immunol 77:102190

Gu J, Lu S (2021) An effective intrusion detection approach using svm with naïve bayes feature embedding. Comput Secur 103:102158

Gujar N, Tambe M, Parande M, Salunke N, Jagdale G, Anderson SG, Dharmadhikari A, Lakhkar A, Kulkarni PS (2021) A case control study to assess effectiveness of measles containing vaccines in preventing severe acute respiratory syndrome coronavirus 2 (sars-cov-2) infection in children. Hum Vaccines Immunother 17(10):3316–3321

Hasan SM, Anisha AM, Adnin R, Eliza IJ, Tarin I, Afroz S, Islam AAA (2022) Revealing influences of socioeconomic factors over disease outbreaks. In: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS), pp 490–512

Hayman DT (2019) Measles vaccination in an increasingly immunized and developed world. Hum Vaccines Immunother 15(1):28–33

Jijo BT, Abdulazeez AM (2021) Classification based on decision tree algorithm for machine learning. Evaluation 6:7

Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, Brugnara G, Schell M, Kessler T, Foltyn M et al (2019) Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol 20(5):728–740

Kissi J, Owusu-Marfo J, Osei E, Dzamvivie K, Akorfa Anku V, Lamptey Naa Lmiokor J (2022) Effects of coronavirus pandemic on expanded program on immunization in weija gbawe municipality (accra-ghana). Hum Vaccines Immunother 18(6):2129830

Kuehn BM (2021) Drop in vaccination causes surge in global measles cases, deaths. Jama 325(3):213–213

Kumar S, Mallik A, Kumar A, Del Ser J, Yang G (2023) Fuzz-clustnet: Coupled fuzzy clustering and deep neural networks for arrhythmia detection from ecg signals. Comput Biol Med 153:106511

Liu S, Lu M, Li H, Zuo Y (2019) Prediction of gene expression patterns with generalized linear regression model. Front Genet 10:120

Malhotra R, Kamal S (2019) An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. Neurocomputing 343:120–140

Messina N, Zimmermann P, Curtis N (2019) The impact of vaccines on heterologous adaptive immunity. Clin Microbiol Infect 25(12):1484–1493

Mirzaei G, Adeli H (2022) Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia. Biomed Signal Process Control 72:103293

Misin A, Antonello RM, Di Bella S, Campisciano G, Zanotta N, Giacobbe DR, Comar M, Luzzati R (2020) Measles: an overview of a re-emerging disease in children and immunocompromised patients. Microorganisms 8(2):276

Muhammad L, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA (2021) Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. SN Comput Sci 2:1–13

Patel MK, Goodson JL, Alexander JP Jr, Kretsinger K, Sodha SV, Steulet C, Gacic-Dobo M, Rota PA, McFarland J, Menning L et al (2020) Progress toward regional measles elimination-worldwide, 2000–2019. Morbidity Mortality Wkly Rep 69(45):1700

Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine Learning, pp 101–121. Elsevier

Portnoy A, Jit M, Ferrari M, Hanson M, Brenzel L, Verguet S (2019) Estimates of case-fatality ratios of measles in low-income and middle-income countries: a systematic review and modelling analysis. Lancet Global Health 7(4):472–481

Rao AS, D'Mello DA, Anand R, Nayak S (2021) Clinical significance of measles and its prediction using data mining techniques: a systematic review. Advances in artificial intelligence and data engineering

Rao AS, BH KP, Nayak S, Shenoy RD (2022) Detection of epithelial giant cells in nasal aspirate cytological smears using deep learning and computer vision techniques: an approach for early diagnosis of measles disease. Philipp J Sci 151(6A):2129–2143

Ristić M, Milošević V, Medić S, Djekić Malbaša J, Rajčević S, Boban J, Petrović V (2019) Sero-epidemiological study in prediction of the risk groups for measles outbreaks in Vojvodina, Serbia. PLoS One 14(5):0216219

Saladi S, Karuna Y, Koppu S, Reddy GR, Mohan S, Mallik S, Qin H (2023) Segmentation and analysis emphasizing neonatal mri brain images using machine learning techniques. Mathematics 11(2):285

Sanson G, Welton J, Vellone E, Cocchieri A, Maurici M, Zega M, Alvaro R, D'Agostino F (2019) Enhancing the performance of predictive models for hospital mortality by adding nursing data. Int J Med Inform 125:79–85

Sanyaolu A, Okorie C, Marinkovic A, Ayodele O, Abbasi AF, Prakash S, Gosse J, Younis S, Mangat J, Chan H (2019) Measles outbreak in unvaccinated and partially vaccinated children and adults in the united states and canada (2018–2019): a narrative review of cases. INQUIRY J Health Care Organ Provis Financ 56:0046958019894098

Sharma N, Raghava, GPS (2022) Computational tools for designing therapeutic molecules against virulent factors of pathogens. PhD thesis, IIIT-Delhi

Shen W, Ye H, Zhang X, Huo L, Shen J, Zhu L, Wang X, Cui D (2020) Elevated expansion of follicular helper t cells in peripheral blood from children with acute measles infection. BMC immunol 21(1):1–8

Stephenson J (2021) Measles a growing global threat as Covid-19 disrupts childhood vaccinations. JAMA Health Forum 2:214680–214680

Wang S, Yao X (2012) Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybernet Part B (Cybernetics) 42(4):1119–1130

Xu W, Zhang J, Zhang Q, Wei X (2017) Risk prediction of type ii diabetes based on random forest model. In: 2017 third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB), pp 382–386. IEEE

Yadav A (2021) Predicting covid-19 using random forest machine learning algorithm. In: 2021 12th international conference on computing communication and networking technologies (ICCCNT), pp 1–6. IEEE

Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, Zheng S, Xu A, Lyu J (2020) Brief introduction of medical database and data mining technology in big data era. J Evid Based Med 13(1):57–69

Zhu T, Lin Y, Liu Y (2017) Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognit 72:327–340

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.