Health and Technology
https://doi.org/10.1007/s12553-021-00551-9

ORIGINAL PAPER

On the goodness of fit of parametric and non-parametric data mining techniques: the case of malaria incidence thresholds in Uganda

Francis Fuller Bbosa[1,2] · **Josephine Nabukenya**[2] · **Peter Nabende**[2] · **Ronald Wesonga**[3]

Received: 22 December 2020 / Accepted: 14 April 2021

Abstract
To identify which data mining technique (parametric or non-parametric) best fits the predictions on imbalanced malaria incidence dataset. The researchers compared parametric techniques in form of naïve Bayes and logistic regression against non-parametric techniques in form of support vector machines and artificial neural networks and their goodness of fit and prediction was assessed using 10-fold and 5-fold cross-validation on an independent validation dataset set to determine which model best fits the predictions on imbalanced malaria incidence dataset. The 10-fold cross-validation outperformed the 5-fold cross-validation in all performance metrics with the naïve Bayes classifier attaining accuracy of 69% with a sensitivity of 90.9%, a specificity of 55.6%, a precision of 55.6% and F-measure score of 69.0%, the logistic regression achieved an accuracy of 65.5% with a sensitivity of 83.3%, a specificity of 52.9%, a precision of 55.6% and F-measure score of 66.7%, the support vector machines achieved an accuracy of 82.8% with a sensitivity of 88.2%, a specificity of 75.0%, a precision of 83.3%, and F-measure score of 85.7% whereas the artificial neural networks registered an accuracy of 89.7% with a sensitivity of 94.1%, a specificity of 83.3%, a precision of 88.9%, and F-measure score of 91.4%. Non-parametric data mining techniques in form of artificial neural networks and support vector machines outperformed the parametric data mining technique in form of naïve Bayes in making predictions emanating from imbalanced malaria incidence dataset on account of registering higher F-measure values of 91.4% and 85.7% respectively.

Keywords Data mining · Prediction · Parametric · Non-parametric · Comparison · Malaria

1 Introduction

In the past decade, machine learning models particularly data mining have gained the attention of several scholars [1–4] while undertaking predictive studies. According to Hagenauer, Omrani and Helbich [5], data mining encompasses several inductive techniques that identify hidden patterns, by repetitively learning from training data and relating a target output attribute to underlying explanatory attributes. The learned model from the training data can then be used to classify or predict previously unknown instances [6, 7]. Agyapong, Hayfron-Acquah, & Asante [8] assert that predictive data mining approaches also known as classification learns from the training set, where all attributes are already associated with known class labels and build a model which is used to estimate unknown values of new attributes [9, 10].

Furthermore, predictive data mining techniques are split into parametric and non-parametric depending on the nature of assumptions about the form of relationship between the antecedent and consequent attributes [11, 12]. Parametric techniques in the context of machine learning assume a finite set of parameters and underlying assumptions about data structure whereas non-parametric are generalized since they do not take into consideration any assumptions about the probability distribution of the data [13]. Parametric data mining techniques such as linear or multi regression, and naïve Bayes have gained popularity as predictive and heuristic models [11] due to their capability to comprehend underlying interactions among attributes in data. Whilst these parametric models have conventionally contributed to understanding underlying relationships and assumptions

✉ Francis Fuller Bbosa
fullerbbosa@gmail.com

1 School of Statistics and Planning, Makerere University, Kampala, Uganda

2 School of Computing and Informatics Technology, Makerere University, Kampala, Uganda

3 Department of Statistics, College of Science, Sultan Qaboos University, Muscat, Oman

Published online: 21 April 2021

between antecedent and consequent attributes in predictive modeling [14, 15], their results are only reliable when model assumptions are fulfilled [5].

On the otherhand, a number of non-parametric data mining techniques have been developed to circumvent problems associated with parametric predictive techniques [16, 17]. Some of these include Decision Trees [18], Support Vector Machines (SVM) [18], Artificial Neural Networks [19], K Nearest Neighbour [20]. In contrast to parametric techniques, non-parametric techniques depend on machines (computers) to examine the data for its structure, devoid of the concept of the underlying data structure assumptions [17]; henceforth permitting dynamic data structure for modeling, which lead to improved predictive abilities [21, 22]. Nevertheless, the performance of estimates emanating from parametric and non-parametric predictive data mining techniques have not been compared in a systematic manner particularly on similar imbalanced data and thus being a subject of debate [23–26]; which presents a challenge of which appropriate model to adopt with regards to imbalanced dataset. Imbalanced data refers to a classification dataset where the number of instances in the consequent attribute are not uniformly represented [27–29] wherever the majority of instances are recorded for one of the classes (majority class) and fewer instances for the other class(es) [30]. Consequently, the ultimate alarm raised by the imbalanced learning obstacle is that the performance of standard predictive data mining techniques is significantly compromised due to failure to take into consideration class imbalance; leading to a reduction in performance of the classifier as a result of inaccurate estimates [27, 30, 31]. Pertinent literature surveyed suggests that prior to undertaking data mining, the imbalanced data situations can be resolved by approaches such as resampling in form of under-sampling to eliminate selected observations of the dominant class and over-sampling to generate fresh observations of the lesser class respectively [27, 32, 33] and hybridization in form of applying boosting and bagging based ensemble algorithms [33, 34].

Numerous researchers [2, 35–37] attest that some of the most successful application of predicting abilities of both parametric and non-parametric data mining techniques have been demonstrated in disease surveillance systems. Unfortunately, majority of datasets stemming from disease surveillance systems are often imbalanced [38, 39], where the occurrence of an event in one target class significantly varies from that of other target classes [23, 40]. Thus an imbalanced dataset is a precursor to biasedness and variance in predictive estimates [41], which is a hindrance in building reliable prediction models [39, 40]. Additionally, in the healthcare domain, dataset attributes habitually contain incomplete and heterogeneous instances [42, 43]. Moreover, due to concerns of patient privacy, and the proprietary nature of electronic medical records, the healthcare databases are always imbalanced in nature [43]. This skewness in the distribution of class attributes creates a serious challenge of biasedness in the predictive modeling of several healthcare related datasets [43].

1.1 Malaria as a case study

The problem of imbalanced data is very common in the healthcare industry and has been underscored by several scholars [44–47]. Hence, majority of healthcare datasets such as malaria usually have fewer positive cases, when compared to the number of negative patients in the dataset [48]. In undertaking this study, malaria incidence rates in Uganda were used as the disease model due to the fact that current technological malaria predicting systems are insufficient at estimating the extent of malaria incidence rates, principally in highly endemic countries such as Uganda [49]. This may possibly be due to the presence of imbalanced data, given that it's a key challenge to achieving successful estimates from predictive data mining algorithms [24, 35]. Malaria incidence rate refers to the number of malaria cases per 1,000 population at risk [50]. Malaria incidence rates enable the calculation of incidence thresholds for activating an epidemic or outbreak alert signal [51]. Above all, malaria remains a fundamental health problem in Uganda accounting for atleast 20% of all hospital deaths [52].

Hence, the purpose of this paper is to assess the performance of parametric and non-parametric data mining techniques for predicting malaria incidence thresholds from a similar imbalanced dataset, leading to a suitable data mining technique for predicting future monthly malaria incidence thresholds in Uganda's urban areas. In this study, the researcher compared a parametric technique in form of naïve Bayes classifier [3, 53] and logistic regression [54] against non-parametric techniques in form of SVM [55–57] and ANN [58–60]. This is on account of their greater ability in modeling classification type prediction problems [61] coupled with the fact that all are logic-based systems, which tend to perform better when dealing with categorical consequent attributes [62, 63].

Motivated by the fact that real-life datasets are imbalanced in nature and there is no specific recommended data mining technique that works best under such scenarios; thus creating a challenge of which type of technique to adopt. Hence, the need to evaluate the performance of parametric and non-parametric predictive data mining techniques in a systematic manner using a similar imbalanced dataset.

1.2 Problem statement

A key limitation for parametric data mining techniques is their reliance on several underlying assumptions

particularly linearity, which are too rigid for the predominantly non-linear real-life data modeling investigations, homoscedasticity and absence of correlation among explanatory attributes. Failure to fulfill the above assumptions may lead to inaccurate and unreliable estimates. To address these pitfalls, non-parametric machine learning methods have been proposed on the basis of their ability to make fewer assumptions and proficiency of learning from more data as the number of attributes increases; henceforth permitting a dynamic data structure for predictive modeling, leading to improved estimates. Nevertheless, parametric and non-parametric data mining techniques have not been compared in a systematic manner on the same imbalanced dataset and thus a subject of debate, which creates a challenge of which type of technique to adopt.

Additionary, imbalanced data is very common in the healthcare domain particulary malaria datasets due to the presence of heterogeneous and incompete instances, as well as considerations of patient privacy and the patented nature of electronic medical records. This skewness in the distribution of class attributes generates a key challenge of biasedness in the predictive modeling of malaria incidences. Hence, the need to identify which model (parametric or non-parametric) best fits the predictions on imbalanced malaria incidence dataset.

2 Literature review

A number of studies have been undertaken in the past decade utilizing either parametric or non-parametric data mining techniques amenable to the prediction of malaria incidence rates [64–66].

In 2012, Oluwagbemi & Clarence [66] built predictive models to control and reduce malaria incidences in Africa. Based on an artificial neural network intelligence system, experimentation was executed by feeding the annual data of each country into the predicting system and controlling some malaria occurrence influencing attributes to 30%, 60%, and 90% respectively. At 90% predicting intensity, the system showed that malaria incidences would reduce by 2014 in all the study nations. For instance, malaria incidences were estimated to decline by 15.71% in Madagascar, 38.44% in Nigeria, 38.98% in Sudan, 42.61% in Kenya, and 45.21% in Uganda. They concluded that the generated system could be used to predict future malaria occurrences by governments, and other relevant health agencies for appropriate public health planning.

In 2013, Zacarias and Boström [67] developed a model to predict malaria incidences in rural settings of Mozambique. They used support vector machines and the developed model performed better than other models when employing cross-validation on the training set since it obtained the smallest Mean Square Error (MSE) of 0.65% and it was thus chosen for further analysis. The somewhat small MSE suggests that the developed model is useful for predicting malaria incidences in the surveyed area with acceptable accuracy.

In order to predict malaria incidence using incidence records and weather patterns in Indonesia, Arifianto, Barmawi, and Wibowo [68] employed artificial neural networks. The researchers proposed a modified Neural Network to reduce the learning time and computation while maintaining accuracy in predicting Malaria incidence by relating it to weather patterns. It was proven that the modified GMDH PNN was able to reduce the learning time by 72% and improve the accuracy to 88.02%.

To predict the malaria epidemic in the Indian state of Maharashtra, Sharma et al. [69] employed Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The researchers observed that the SVM was more accurate than ANN since the SVM model predicted the outbreak 15 -20 days in advance. They concluded that the accuracy of estimates can be increased using more training data and scaling it up to the country level. Buczak et al. [70] proposed "Fuzzy association rule mining and classification for the prediction of malaria in South Korea". The authors wanted to extract relationships between epidemiological, meteorological, climatic, and socio-economic data from Korea with respect to predicting malaria using association rules with future malaria cases classified as LOW, MEDIUM, or HIGH. HIGH was considered an outbreak based on user recommendations. The fuzzy association rule technique produced a positive predictive value (PPV) and Sensitivity scores 0.842 and 0.681 respectively, for the HIGH classes. The fuzzy model estimates were considerably better than those obtained by other models generated using a decision tree, Support Vector Machine, Random Forest, and Holt-Winters techniques. Hence, this study proves that a data-driven methodology based on data mining can be utilized for the prediction of different disease incidences.

In their study to test the efficacy of artificial neural networks on malaria abundances, Santosh and Ramesh [71] used clinical and environmental variables collected from Khammam district in India. Study findings revealed that clinical data such as the number of patients treated with symptoms and without symptoms can improve the prediction level when combined with environmental variables. The average error of the prediction model ranged from 18% to 117%. They concluded that more exploration is required in the prediction of malaria using big data to improve the accuracy in real practice.

Olayinka and Chiemeke [4] proposed "Predicting Pediatric Malaria Occurrence Using Classification Algorithm in Data Mining". Using decision tree classification algorithms, a model was developed to predict the occurrence of malaria in children aged less than six (5) years. Findings reveal that the J48 generated better results with the least root

Table 1 Description of variables used in this study

Label	Description	Data Type	Definition
U5_POP2	The population of persons under 5 years of age	Numeric	The average number of people under the age of 5
Max_Temp2	Average maximum temperature	Numeric	The average maximum temperature recorded in a month in degrees Celcius
Min_Temp2	Average minimum temperature	Numeric	The average minimum temperature recorded in a month in degrees Celcius
Rainfall2	Rainfall totals	Numeric	Total rainfall recorded in a month in millimeters
Incidence_ rate_ Threshold	Monthly incidence threshold of malaria per 1,000 population	Categorical	The number of monthly malaria cases per 1,000 population at risk where Moderate ="21-37 cases per 1,000 population at risk" and Low="8-20 cases per 1,000 population at risk".

mean squared error of 0.1641, thus improving the accuracy of the prediction. Hence the generated model is an efficient and effective tool for early detection of malaria incidence in children to diminish the mortality rate.

Generally, this study emphasizes the need to identify which model (parametric or non-parametric) best fits the predictions on imbalanced malaria incidence data given that several researchers have applied dissimilar data mining techniques for generating estimates from the same imbalanced data with varying results. Consequently, the current study compared parametric techniques in form of naïve Bayes and logistic regression classifiers against non-parametric techniques in form of SVM and ANN classifiers on the same imbalanced data in order to identify and recommend an appropriate technique to adopt under scenarios of predictions based on imbalanced consequent attributes.

3 Methods

3.1 Data preprocessing

3.1.1 Data sources

Malaria incidence data were obtained from the Ministry of Health through the District Health Information Software (DHIS2), meteorological data were obtained from Uganda National Meteorological Authority (UNMA)[1], whereas demographic data were obtained from Uganda Bureau of Statistics (UBoS)[2]. Monthly data were extracted for the period January 2012 to December 2019 for Kampala from all the above-stated sources (Table 1).

A total of 96 instances, each instance representing a monthly observation from January 2012 to December 2019 were collected. The attributes in Table 1 were selected due to the fact that evidence from several studies [72, 73] corroborate that malaria incidence in highly endemic countries such as Uganda is closely related to meteorological and demographic conditions particularly among children aged below five years. Additionally, temperature and rainfall were the only meteorological attributes collected consistently by formal government of Uganda entities over the past decade.

3.1.2 Data cleaning

The researchers resolved inconsistencies in data by augmenting missing data through linear interpolations particularly for monthly data between subsequent years using the following equation [74];

$$\left(\frac{z - z_0}{x - x_0}\right) = \left(\frac{z_1 - z_0}{x_1 - x_0}\right) \tag{1}$$

where z_0 and z_1 were the existing data in x_0 and x_1 months respectively. The population parameter z was linearly interpolated in each month.

3.1.3 Data transformation

The researchers employed z-scores to normalize the data as a way of re-scaling all attribute values, thus ensuring that all non-categorical antecedent attribute values were in similar ranges [74, 75]. The mean and standard deviation of the attributes were used for normalization as illustrated in equation (2);

$$D_i = \frac{(b_i - \mu)}{\delta} \tag{2}$$

where D_i=normalised attribute value b_i =original attribute value, μ=mean attribute value and δ=standard deviation of the attribute.

Additionally, the researchers also discretized continuous attributes in order to develop more comprehensible intervals and reveal non-linear interactions among attributes in the dataset [76]. Furthermore, discretization facilitated an improvement in the predictive accuracy of a classifier

[1] www.unma.go.ug

[2] www.ubos.org

through the derivation of discrete data [68]. The procedure adapted from Li and Wang [77] was employed during the discretization process;

i Let B be a continuous attribute with $[x, y]$ as its domain values
ii We establish cut-off thresholds $(a_1, a_2, \ldots\ldots, a_{m-1})$, where $x < a_1 < a_2 < \cdots < a_{m-1} < y$
Hence separating $[x, y]$ into m disjoint thresholds .i.e. $[x, a_1)$, $(a_1, a_2]$,…, $(a_{m-1}, y]$
iii The continuous values of B were transformed into m different discrete values $(d_1, d_2, \ldots\ldots, d_m)$, as illustrated in equation (3);

$$d_m \text{ if } a_{m-1} \leq B \leq y \tag{3}$$

where index $i = 1, 2, 3 \ldots\ldots, \ m-1$

3.1.4 Deriving malaria incidence thresholds

In deriving malaria incidence thresholds, the researchers embraced a World Health Organization (WHO) classification for various transmission settings with high defined as greater than or equal to 450 cases per 1,000 people annual parasite incidence (API); moderate defined as 250 to 450 cases per 1,000 API; low defined as 100 to 250 cases per 1,000 API; and very low defined as less than 100 cases per 1,000 API [78]. The researchers divided the above thresholds by 12 in order to generate the average monthly incidence rate thresholds. This implies that the average monthly incidence rate classification would be defined as:

i High denoting "> = 38 cases per 1,000 population at risk"
ii Moderate denoting "21-37 cases per 1,000 population at risk"
iii Low denoting"8-20 cases per 1,000 population at risk"
iv Very low denoting "< 8 cases per 1,000 population at risk"

3.1.5 Training and validation datasets

The researchers split the dataset into training and testing datasets. 70% of the dataset was assigned to the training group for the development of the classifiers. The rest of the dataset (30% of the total cases) was assigned to the validation groups for the assessment of model performance [79].

3.1.6 Testing classifier assumptions

Assumptions of various classifiers employed in this study were tested before the models were fit on training dataset. In cases were the assumptions such as normality of the dataset were not initially met, the data pre-processing phases as indicated in equations (1-3) were employed to transform the initial dataset to meet the basic requirements for the assumptions.

3.2 Analysis

Predictive models using the following data mining techniques: naïve Bayes; a classifier for computing the posterior probability that an antecedent attribute belongs to a target class [80, 81], logistic regression; a classifier that builds a linear combination of attributes to compute each class value of the target attribute [82–84], support vector machines; a classifier which utilizes a non-linear mapping procedure to transform training data into a higher dimension by maximizing the distance between the closest instances within the class labels of the target attribute [85–87] and artificial neural networks; a classifier with the capability to learn from training data through reiteration without prior knowledge about the relationships between input and output variables [58, 60, 88] were established and learnt based on the training dataset and later evaluated using test dataset using the R programming software.

3.3 Goodness of fit and prediction

The researchers employed a k-fold cross-validation (CV) method and confusion matrix evaluation metrics.

3.3.1 Evaluation metrics

The researchers computed and compared the performance of the naive Bayes, logistic regression, SVM, and ANN classifiers using a confusion matrix. The classifiers' performance was assessed using accuracy, sensitivity, specificity, precision, and F-measure as illustrated in Table 2. Due to the imbalanced nature of the dataset under investigation, the researchers mainly used one performance metric to assess the overall performance of the classifiers; the F-measure, which is the harmonic mean of precision and recall. In general, precision and recall are trade-offs; that is, if the recall is low, then the precision will be high [89]. The researcher chose the F-measure because several studies [90–93] suggest that the F-measure is a more reliable metric among all classification metrics emanating from imbalanced data. Nevertheless, the researchers computed and compared other metrics by the classifier as illustrated in Table 2.

In the context of this study, the entries in the confusion matrix were defined as:

i True positive (TP): is the number of actual "LOW" instances classified as "LOW".

Table 2 Performance metrics computed [94]

Metric	Formula	Description
Accuracy/recognition rate (%)	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	Number of correctly classified malaria incidence thresholds to total number of incidences
Sensitivity/ true positive rate (%)/Recall	$\frac{TP}{(TP+FN)}$	The proportion of low incidence thresholds that are correctly classified
Specificity/ true negative rate (%)	$\frac{TN}{(FP+TN)}$	The proportion of "moderate" incidence thresholds that are correctly classified
Precision (%)	$\frac{TP}{(TP+FP)}$	The proportion of "low" incidences predicted to be "low" that are truly "low" incidences
F-Score/F-measure	$\left(\frac{2 * Precision * Recall}{Precision + Recall}\right)$	The harmonic mean of precision and recall

ii False-positive (FP): is the number of actual "MODERATE" instances classified as "LOW"
iii False Negative (FN): is the number of actual "LOW" instances classified as "LOW".
iv True Negative (TN): is the number of actual "MODERATE" instances classified as "MODERATE".

3.3.2 Classifier validation and generalizability

Ascertaining the validity and generalizability of a model helps determine how well it can classify new participants who may have dissimilar characteristics than those in the original sample [95]. The researchers employed the k-fold cross-validation approach [96] where the data was randomly partitioned into k disconnected groups, and one group at a time was used for classifier testing while the remaining k-1 groups were used for classifier training [97].

The algorithm adapted from [98] is as follows:

i Randomly divide the data set into K groups (K-fold)
ii Reserve 1 group and train the classifier on all the remaining (K-1) groups
iii Test the classifier on the reserved group and register the prediction error
iv Repeat this process until each of the k groups has served as the test set.
v Compute the average of the K registered errors, which serves as the performance metric for the classifier.

For instance, the researchers mathematically computed the 10 k cross-validation based on equation (4);

$$10\text{ k cross validation} = \frac{1}{N}\sum_{k=1}^{10}\sum_{igroup_k}\left(y_i - g_i^k\right) \quad (4)$$

where g_i^k denotes the predicted value of y_i from the model estimated on all instances not found in $group_k$.

Additionally, in order to generalize the performance of the classifiers on new data, we assessed classifier performance on test data, which was not originally included in the training estimation [99].

3.4 Software

The researchers undertook data processing and analysis entirely in R, version 3.6.3 [100], by means of R packages "funModeling" version 1.9.3 [101], "dplyr" version 0.8.5 [102], "tidyr" version 1.0.2 [103], "caret" version 6.0.86 [104], packages.

4 Results

In this section, the researchers first present the results from each classifier and then presents the comparison results.

4.1 Malaria incidence thresholds

Findings revealed that the "High" and "Very low" thresholds were none existent in malaria cases recorded for Kampala for the study period under investigation and thus, the target attribute eventually comprised two classes ("Low" and "Moderate") as the thresholds for malaria incidence rates. Therefore, for the period under consideration, malaria cases were estimated to be as low as 8 cases per 1000 population at risk and not exceeding 37 cases per 1000 population at risk. With regards to imbalancedness, the ratio of imbalancedness for the malaria incidence thresholds was 0.54:0.46 in favour of the "low" class.

4.2 Comparison of the 5-fold and 10-fold cross-validation

The prediction results of each model for test data were calculated over both five and ten random samples generated by the 5-fold and 10-fold cross-validation procedures respectively. The researchers evaluated the performance of the classification algorithms using a confusion matrix, which revealed actual versus predicted values as illustrated in Table 3.

Table 3 Comparison of the 10-fold and 5-fold cross-validation with their confusion matrix metrics

Classifier	K-fold Cross-Validation	TP	FN	FP	TN
Naïve Bayes	5-fold	9	1	9	10
	10-fold	10	1	8	10
Logistic regression	5-fold	10	2	8	9
	10-fold	10	2	8	9
SVM	5-fold	13	1	5	10
	10-fold	15	2	3	9
ANN	5-fold	15	4	3	7
	10-fold	16	1	2	10

From Table 3, a comparison of the diagonal metrics totals on the confusion matrices for both 5 and 10-fold cross-validation revealed that the 10-fold was a better performer across all classifiers, and thus the researchers focussed on results emanating from the 10-fold cross-validation confusion matrix for the subsequent analysis.

4.3 The 10-fold cross-validation prediction results

Using 10-fold cross-validation performance metrics from Table 3, the performance of the classifiers was evaluated based on their capacity to classify the instances of the data set into "Low" and "Moderate" malaria incidence thresholds. Calculation of the performance metrics indicated in Table 3 revealed the following results indicated in Fig. 1.

Figure 1 shows the accuracy, sensitivity, specificity, precision, and F-Score of each consequent class attained by the four techniques. The results reveal that the parametric naïve Bayes and logistic regression models correctly classified (sensitivity) 90.9% and 83.3% of malaria cases in the low incidence threshold and only 55.6% and 52.9% of the cases in the moderate incidence threshold respectively. On the other hand, the non-parametric SVM and ANN model correctly classified 88.2% and 94.1% of malaria incidence rates in the low incidence threshold, as well as 75% and 83.3% of malaria incidence rates in the moderate incidence threshold respectively.

Additionally, both naïve Bayes and logistic regression classified only 55.6% of the low incidences correctly (precision) whereas the SVM, and ANN classified 83.3%, and 88.9% of the low incidences correctly (precision) respectively.

From the observations made in Fig. 1, the non-parametric classifiers achieved higher prediction accuracy for incidence thresholds at 89.7% and 82.8% for the ANN and SVM respectively compared to 69% and 65.5% attained by the parametric naïve Bayes and logistic regression models respectively. Furthermore, the non-parametric SVM and ANN were more specific since they identified 75% and 83.3% of the "negative" moderate incidences respectively compared to 55.6% and 52.9% registered by the parametric naïve Bayes and logistic regression classifiers.

Nonetheless, the results in Table 1 show that the consequent attribute classes were imbalanced and therefore the F-measure is a more reliable metric than accuracy is such circumstances [91, 92]. Subsequently, the best classifier

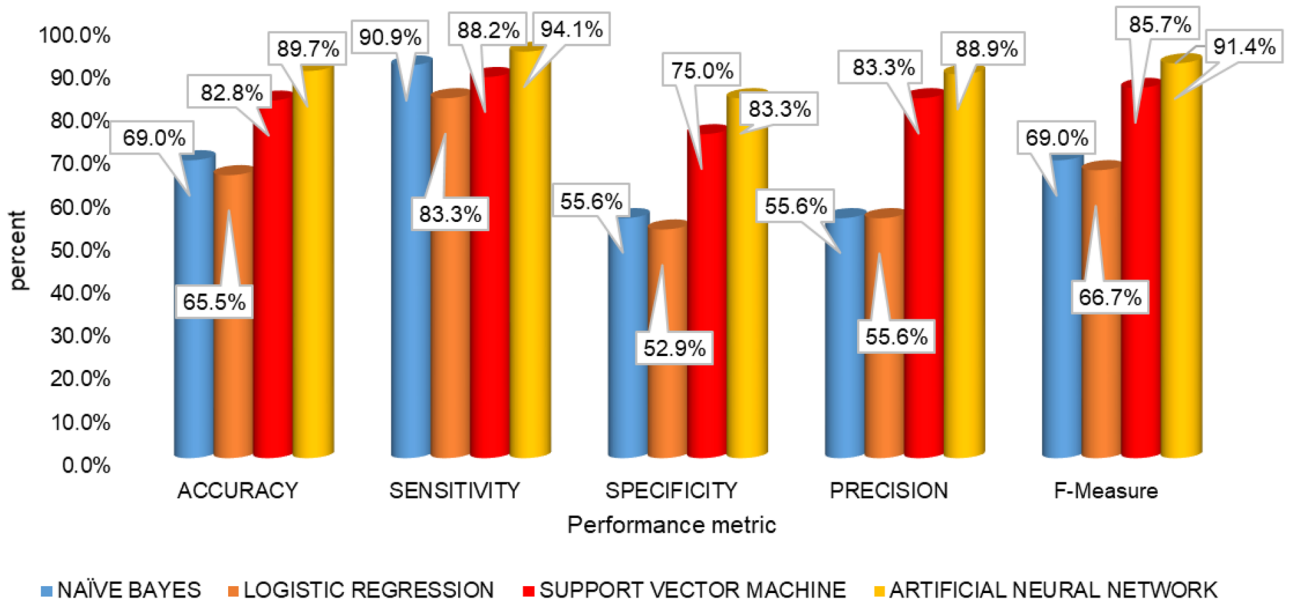

Fig. 1 Comparison of classifiers' performance using 10 fold cross-validation. NB: Microsoft Excel was used to generate the figures

in scenarios of imbalanced data is the classifier with the highest F-measure value irrespective of its accuracy metric and thus the non-parametric techniques in form of ANN and SVM were identified as better classifiers than the parametric technique in form of naïve bayes on the account of registering a higher F-measure value of 91.4% and 85.7% respectively.

5 Discussion

The main aim of this research is to evaluate the performance of parametric and non-parametric predictive data mining techniques in a systematic manner using a similar imbalanced dataset. The researchers compared parametric and non-parametric classifiers for predicting malaria incidences thresholds emanating from an imbalanced dataset. The specific techniques employed were naïvebayes, logistic regression, SVM, and ANN respectively. The results showed that the non-parametric models in form of SVM and ANN outperformed the parametric model in form of naïve Bayes and logistic regression in terms of accuracy, specificity, precision, and F-measure whereas the naïve Bayes outperformed the SVM and only in terms of the sensitivity metric on imbalanced data.

Additionally, the findings are in agreement with the notion that high sensitivity and specificity may not be attainable in real-world situations concurrently [105] due to the fact that they are inversely related, implying that as the specificity increases, the sensitivity decreases and vice versa [106]. Hence there is a trade-off between sensitivity and specificity for both parametric and non-parametric models with the SVM and ANN classifiers recording a lower specificity of 75% and 83.3% and higher sensitivity of 88.2% and 94.1% respectively. A similar trend was observed among the parametric models with the naïve Bayes and logistic regression registering a higher sensitivity of 90.9% and 83.3% compared to a lower specificity of 55.6% and 52.9% respectively.

On the other hand, the findings of this study are in agreement with those of other previous studies [94] [107] that suggest that the nature of underlying data assumptions such as imbalanced target attributes significantly influence the success of a predictive data mining technique. A key limitation of this study is that the researchers did not take into consideration the effect of merging parametric and non-parametric techniques in order to utilize the strengths of each individual technique and compensate for each other's weaknesses [108–110].

6 Conclusion

The researchers compared parametric classifiers in the form of naive Bayes and logistic regression against non-parametric classifiers in the form of SVM and ANN models to determine which model best fits the predictions on imbalanced data. Findings revealed that the naïve Bayes classifier attained an accuracy of 69% with a sensitivity of 90.9%, a specificity of 55.6%, a precision of 55.6% and F-measure score of 69%, the logistic regression achieved an accuracy of 65.5% with a sensitivity of 83.3%, a specificity of 52.9%, a precision of 55.6% and F-measure score of 66.7%, the SVM achieved an accuracy of 82.8% with a sensitivity of 88.2%, a specificity of 75%, a precision of 83.3% and F-measure score of 85.7% whereas the ANN registered an accuracy of 89.7% with a sensitivity of 94.1%, a specificity of 83.3%, a precision of 88.9% and F-measure score of 91.4%. However, on account of the imbalanced nature of the consequent attribute, the best-classifier identification was made based on the F-measure metric instead of accuracy. Hence after comparative analysis, we concluded that non-parametric data mining techniques in form of ANN and SVM outperformed the parametric data mining techniques in form of naïve Bayes and logistic regression in making predictions emanating from imbalanced data because they registered a higher F-measure metric of 91.4% and 85.7% respectively.

Above all, the current study could benefit future works, particularly the amalgamation of parametric and non-parametric data mining techniques that would address the drawbacks of the independent parametric and non-parametric techniques which the researchers intend to investigate in the near future.

Acknowledgements The authors extend their appreciation to Mr. Douglas Candia and Mr. Frank Namugera who contributed to improving this research. This research was partly funded by Makerere University through the Staff Development, Welfare and Retirement Benefits Committee (SDWRBC).

Authors' contributions FFB was involved in drafting the proposal, data collection, data preprocessing, data analysis, model designing and writing the manuscript. JN, PN and RW were supervisors of the work. All authors read amd approved the final manuscript.

Funding This study was partly funded by Makerere University through the Staff Development, Welfare and Retirement Benefits Committee (SDWRBC).

Availability of data and material The data was sourced from the ministry of health(www.health.go.ug), Uganda Bureau of Statistics (www.ubos.org) and Uganda National Meteorological Authority (www.unma.go.ug)and it has been availed/uploaded as supplementary material.

Code availability The program scripts/code can be availed by the first author upon request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Ferreira D, Oliveira A, & Freitas A. Applying data mining techniques to improve diagnosis in neonatal jaundice. In Med Inform Decis Mak. 2012;12(143):2–7.
2. Hakizimana L, Cheruiyot K, Kimani S, Nyararai M. A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets. Int J Comput. (IJC). 2017;27(1):69–83.
3. Kotlar AM, Jong De, van Lier Q. Evaluation of parametric and nonparametric machine-learning techniques for prediction of saturated and near-saturated hydraulic conductivity. Vadose Zone J. 2019. https://doi.org/10.2136/vzj2018.07.0141.
4. Olayinka TC, Chiemeke SC. Predicting paediatric malaria occurrence using classification algorithm in data mining. J Adv Math Comput Sci. 2019;31(4):1–10. https://doi.org/10.9734/JAMCS/2019/v31i430118.
5. Hagenauer J, Omrani H, Helbich M. Assessing the performance of 38 machine learning models : the case of land consumption rates in Bavaria, Germany. Int J Geogr Inf Sci. 2019;1–21. https://doi.org/10.1080/13658816.2019.1579333.
6. Maxwell AE, Warner TA, Fang F. Implementation of machine-learning classification in remote sensing: an applied review. Int J Remote Sens. 2018;39:2784–817.
7. Tayyebi A, Pijanowski BC. Modeling multiple land use changes using ANN, CART and MARS: comparing tradeoffs in goodness of fit and explanatory power of data mining tools. J Appl Earth Obs Geoinf. 2014;28:102–16.
8. Agyapong KB, Hayfron-Acquah J, Asante M. An overview of data mining models (descriptive and predicitve). International Journal of Software & Hardware Research in Engineering. 2016;4(5):53–60. https://doi.org/10.1007/978-3-319-13084-2_59.
9. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int J Comput Sci Appl. 2013;6(2).
10. Krishnaiah V, Narsimha G, Subhash C. Diagnosis of lung cancer prediction system using data mining classification techniques. (IJCSIT) Int J Comput Sci Inf Technol. 2013;4(1):39–45.
11. Goltsman K. Data Mining: Models and Methods. 2017.https://datascience.foundation/sciencewhitepaper/data-mining:-models-and-methods.
12. Ouyang F, Guo B, Ouyang L, Liu Z, Lin S, Meng W. Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. Eur J Radiol. 2019;113(1):251–7. https://doi.org/10.1016/j.ejrad.2019.02.029.
13. Mircioiu C, Atkinson J. A comparison of parametric and non-Parametric methods applied to a Likert Scale. Pharmacy. 2017;5(26):1–12. https://doi.org/10.3390/pharmacy5020026.
14. Abdalrada AS, Yahya OH, Alaidi AHM, Hussein NA, Alrikabi HT, Al-Quraishi T. A predictive model for liver disease progression based on logistic regression algorithm. Period Eng Nat Sci. 2019;7(3):1255–64.
15. David M. Automobile insurance pricing with generalized linear models. Proceedings in GV-Global Virtual Conference (No. 1). 2015.
16. Loucoubar C, Paul R, Bar-hen A, Huret A, Tall A, Sokhna C, Trape J-F, Ly Badara A, Faye J, Diop A, Sakuntabhai A. An exhaustive, non-euclidean, non-parametric data mining tool for unraveling the complexity of biological systems – novel insights into malaria. PLoS One. 2011;6(9):1–16. https://doi.org/10.1371/journal.pone.0024085.
17. Zhao X, Yan X, Yu A, Van Hentenryck P. Prediction and behavioral analysis of travel mode choice : A comparison of machine learning and logit models. Travel Behav Soc. 2020;20:22–35. https://doi.org/10.1016/j.tbs.2020.02.003.
18. Uddin S, Khan A, Hossain ME, Moni MA. (2019). Comparing different supervised machine learning algorithms for disease prediction. In BMC Med Inform Decis Mak. 2019;19(281):1-16. https://doi.org/10.1186/s12911-019-1004-8.
19. Tang Y, Ji J, Gao S, Dai H, Yu Y, Todo Y. A pruning neural network model in credit classification analysis. In Comput Math Methods Med. 2018;(pp. 21–22).
20. Medjahed S, Saadi T, Benyettou A. A Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. Int J Comput Appl. 2013;62(1).
21. Kalaiselvan C, Rao LB. Comparison of reliability techniques of parametric and non- parametric method. Int J Eng Sci Technol. 2016;19:691–9. https://doi.org/10.1016/j.jestch.2015.11.002.
22. Park S, Lee J, Son Y. Predicting market impact costs using nonparametric machine learning models. PLoS Negl Trop Dis. 2016;11(2):1–13. https://doi.org/10.1371/journal.pone.0150243.
23. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. 2011;11(51).
24. Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. In Artificial Intelligence In Medicine: Elsevier B.V; 2019. https://doi.org/10.1016/j.artmed.2019.101723.
25. Norinder U, Boyer S. Binary classification of imbalanced datasets using conformal prediction. J Mol Graph Model. 2017. https://doi.org/10.1016/j.jmgm.2017.01.008.
26. Sambasivam G, Opiyo GD. A predictive machine learning application in agriculture : Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. In Egyptian Informatics Journal: Faculty of Computers and Information, Cairo University; 2020. https://doi.org/10.1016/j.eij.2020.02.007.
27. Mujali OR, López G, Garach L. Bayes classifiers for imbalanced traffic accidents datasets. Accid Anal Prev. 2016;88:37–51. https://doi.org/10.1016/j.aap.2015.12.003.
28. Sarkar BK. Improving disease diagnosis by a new hybrid model. In New Horizons in Translational Medicine 2017;4(1-4):2. Elsevier Ltd. https://doi.org/10.1016/j.nhtm.2017.07.001.
29. Shanab AA, Khoshgoftaar TM, Wald R, Van Hulse J. Comparison of approaches to alleviate problems with high-dimensional and class-imbalanced data. IEEE. 2011;234–239.
30. Wang Z. Practical tips for class imbalance in binary classification. 2018.https://towardsdatascience.com/practical-tips-for-class-imbalance-in-binary-classification-6ee29bcdb8a7.
31. Thammasiri D, Delen D, Meesad P, Kasap N. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst Appl. 2014;41:321–30.
32. Bhatnagar R. Machine Learning and Big Data Processing: A Technological Perspective and Review (Hassanien (ed.). 2018. Springer International Publishing.

33. Krawczyk B. Learning from imbalanced data : open challenges and future directions. Prog Artif Intell. 2016;5:221–32. https://doi.org/10.1007/s13748-016-0094-0.
34. Sun Z, Song Q, Zhu X, Sun H, Xu B, Zhou Y. A Novel Ensemble Method for Classifying Imbalanced Data. In Pattern Recognition: Elsevier; 2014. https://doi.org/10.1016/j.patcog.2014.11.014.
35. Lourenço C, Tatem AJ, Atkinson PM, Cohen JM, Pindolia D, Bhavnani D, Le Menach A. Strengthening surveillance systems for malaria elimination: A global landscaping of system performance, 2015–2017. Malar J. 2019;18(315):1–11. https://doi.org/10.1186/s12936-019-2960-2.
36. Mpimbaza A, Miles M, Sserwanga A, Kigozi R, Wanzira H, Rubahika D, Nasr S, Kapella BK, Yoon SS, Chang M, Yeka A, Staedke SG, Kamya MR, Dorsey G. Short Report: Comparison of routine health management information system versus enhanced inpatient malaria surveillance for estimating the burden of malaria among children admitted to four hospitals in Uganda. Am J Trop Med Hyg. 2015;92(1):18–21. https://doi.org/10.4269/ajtmh.14-0284.
37. Parveen R, Jalbani AH, Shaikh M, Memon KH, Siraj S, Nabi M, Lakho S. Prediction of Malaria using Artificial Neural Network. Int J Comput Sci Netw Secur. 2017;17(12):79–86.
38. Branco P, Torgo L, Ribeiro RP. A Survey of Predictive Modelling under Imbalanced Distributions. 2015.
39. Jain S, Kotsampasakou E, Ecker GF. Comparing the performance of meta-classifiers — a case study on selected imbalanced data sets relevant for prediction of liver toxicity. J Comput Aided Mol Des. 2018;32:583–90. https://doi.org/10.1007/s10822-018-0116-z.
40. Barros TM, Plácido SN, Guedes LA, Silva I. Predictive Models for Imbalanced Data : A School Dropout Perspective. Educ Sci. 2019;9(275). https://doi.org/10.3390/educsci9040275.
41. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high - class imbalance in big data. J Big Data. 2018;5(42). https://doi.org/10.1186/s40537-018-0151-6.
42. Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M. A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data : A Case Study for Brain Tumor Diagnosis. IEEE Access. 2017;4. https://doi.org/10.1109/ACCESS.2016.2647238.
43. Razzaghi T, Roderick O, Marko N, Safro I. Fast imbalanced classification of healthcare data with missing values. 18th International Conference on Information Fusion, 2015;774–781. Washington, DC.
44. Amer AYA, Vranken J, Wouters F, Mesotten D, Vandervoort P, Storms V, Aerts JM. Feature engineering for ICU mortality prediction based on hourly to bi-hourly measurements. Appl Sci. 2019;9(3525). https://doi.org/10.3390/app9173525.
45. González J, Martín F, Sánchez M, Sánchez F, Moreno MN. Multiclassifier systems for predicting neurological outcome of patients with severe trauma and polytrauma in intensive care units. J Med Syst. 2017;41(136). https://doi.org/10.1007/s10916-017-0789-1.
46. Sanchez-Hernandez F, Ballesteros-Herraez J, Kraeim M, Sanchez-Barba M, Moreno-Garcia M. Predictive Modeling of ICU Healthcare-Associated Infections from Imbalanced Data . Using Ensembles and a Clustering-Based Undersampling Approach. Appl Sci. 2019;9(5287). https://doi.org/10.3390/app9245287.
47. Basha HS, Tharwat A, Abdalla A, Hassanien AE. Neutrosophic rule-based prediction system for toxicity effects assessment of biotransformed hepatic drugs. Expert Syst Appl. 2019;121:142–57. https://doi.org/10.1016/j.eswa.2018.12.014.
48. Rao RR, Makkithaya K. Learning from a Class Imbalanced Public Health Dataset : a Cost-based Comparison of Classifier Performance. Int J Electr Comput Eng. 2017;7(4):2215–2222. https://doi.org/10.11591/ijece.v7i4.pp2215-2222.
49. Brown B, Przybylski AA, Manescu P, Caccioli F, Oyinloye G, Elmi M, Al E. Data-Driven Malaria Prevalence Prediction in Large Densely-Populated Urban Holoendemic sub-Saharan West Africa: Harnessing Machine Learning Approaches and 22-years of Prospectively Collected Data. Cornell University. 2019. https://doi.org/10.18907/jjsre.10.Special_105_4.
50. World Health Organization [WHO]. World Malaria Report 2019. 2019.https://www.who.int/publications-detail/world-malaria-report-2019.
51. Wang R, Jiang Y, Michael E, Zhao G. How to select a proper early warning threshold to detect infectious disease outbreaks based on the China infectious disease automated alert and response system (CIDARS). In BMC Public Health 2017;17:1–10. https://doi.org/10.1186/s12889-017-4488-0.
52. Ministry of Health [MoH]. The Uganda malaria reduction strategic plan 2014-2020. Government of Uganda [GoU]. 2014. Retrieved from http://health.go.ug/sites/default/files/TheUgandaMalariaReductionStrategicPlan2014-2020.pdf.
53. Dastile X, Celik T, Potsane M. Statistical and machine learning models in credit scoring: A systematic literature survey. Appl Soft Comput. 2020. https://doi.org/10.1016/j.asoc.2020.106263.
54. Garcia-montemayor V, Martin-malo A, Barbieri C, Bellocchio F, Soriano S, Pendon-ruiz de Mier V, Molina I, Aljama P, Rodriguez M. (2020). Predicting mortality in hemodialysis patients using machine learning analysis. Clin Kidney J. 2020;1–8. https://doi.org/10.1093/ckj/sfaa126.
55. Cui S, Wang D, Wang Y, Yu P, Jin Y. An improved support vector machine-based diabetic readmission prediction. Comput Methods Programs Biomed. 2018;166:123–35. https://doi.org/10.1016/j.cmpb.2018.10.012.
56. Guo X, Li D, Zhang A. Improved support vector machine oil price forecast model based on genetic algorithm optimization parameters. Conference on Computational Intelligence and Bioinformatics. 2012;1:525–30. https://doi.org/10.1016/j.aasri.2012.06.082.
57. Shao Y, Lunetta RS. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS J Photogramm Remote Sens. 2012;70:78–87. https://doi.org/10.1016/j.isprsjprs.2012.04.001.
58. Gao S, Zhao H, Bai Z, Han B, Xu J, Zhao R, Zhang N, Chen L, Lei X, Shi W, Zhang L, Li P, Yu H. Combined use of principal component analysis and artificial neural network approach to improve estimates of PM 2 . 5 personal exposure : A case study on older adults. Sci Total Environ. 2020;726. https://doi.org/10.1016/j.scitotenv.2020.138533.
59. Ragmani A, Elomri A, Abghour N, Moussaid K, Rida M, Badidi E. Adaptive fault-tolerant model for improving cloud computing performance using artificial neural network. Proc Comput Sci. 2020;170:929–34.
60. Yang J, Huang Y, Xu H, Gu D, Xu F, Tang J, Fang C. Optimization of fungi co-fermentation for improving anthraquinone contents and antioxidant activity using artificial neural networks. Food Chem. 2020;313. https://doi.org/10.1016/j.foodchem.2019.126138.
61. Şen B, Uçar E, Delen D. Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Syst Appl. 2012;39(10):9468–76. https://doi.org/10.1016/j.eswa.2012.02.112.
62. Hamblin D, Wang D, Chen G. (2016). Measurement classification using hybrid weighted Naive Bayes. IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA

2016 - Proceedings. 2016.https://doi.org/10.1109/CIVEMSA.2016.7524248.
63. Tamaddoni-nezhad A, Milani GA, Raybould A, Muggleton S, Bohan DA. Construction and Validation of Food Webs Using Logic-Based Machine Learning and Text Mining. In Int Adv Econ Res. 2013;49(1):225–289. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-420002-9.00004-4.
64. Ayo E, Wanjoya A, Luboobi L. Statistical Modeling of Malaria Incidences in Apac District, Uganda. Open J Stat. 2017;7:901–19. https://doi.org/10.4236/ojs.2017.76063.
65. Boruah I, Kakoty S. Analytical Study of Data Mining Applications in Malaria Prediction and Diagnosis. Int J Comput Sci Mob Comput (IJCSMC). 2019;8(3):275–84.
66. Oluwagbemi O, Clarence S. Computational Predictive Framework towards the Control and Reduction of Malaria incidences in Africa. Egypt Comput Sci J. 2012;36(2):1–17.
67. Zacarias O, Boström H. (Predicting the Incidence of Malaria Cases in Mozambique Using Regression Trees and Forests. Int J Electron Comput Sci Eng. (IJCSEE). 2013;1(1).
68. Arifianto A, Barmawi AM, Wibowo AT. Malaria incidence forecasting from incidence record and weather pattern using polynomial neural network. Int J Future Comput Commun. 2014;3(1):60–5. https://doi.org/10.7763/ijfcc.2014.v3.268.
69. Sharma V, Kumar A, Panat L, Karajkhede G, Lele A. Malaria Outbreak Prediction Model Using Machine Learning. Int J Adv Res Comput Eng Technol (IJARCET). 2015;4(12):4415–9.
70. Buczak AL, Baugher B, Guven E, Ramac-Thomas LC, Elbert Y, Babin SM, Lewis SH. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med Inform Decis Mak. 2015;15(1):1–17. https://doi.org/10.1186/s12911-015-0170-6.
71. Santosh T, Ramesh D. Artificial neural network based prediction of malaria abundances using bidata : A knowledge capturing approach. Clinical Epidemiology and Global Health. 2019;7:121–6. https://doi.org/10.1016/j.cegh.2018.03.001.
72. Ssempiira J, Nambuusi B, Kissa J, Agaba B, Makumbi F, Kasasa S, Vounatsou P. Geostatistical modelling of malaria indicator survey data to assess the effects of interventions on the geographical distribution of malaria prevalence in children less than 5 years in Uganda. PLoS One. 2017;12(4):1–20.
73. Texier G, Machault V, Barragti M, Boutin JP, Rogier C. Environmental determinant of malaria cases among travellers. Malar J. 2013;12(1), 1–11. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=23496931.
74. Aggarwal C. Data mining: The Text book. Springer. 2015. https://doi.org/10.1007/978-3-319-14142-814.
75. Crone SF, Lessmann S, Stahlbock R. The impact of preprocessing on data mining : An evaluation of classifier sensitivity in direct marketing. Eur J Oper Res. 2006;173:781–800. https://doi.org/10.1016/j.ejor.2005.07.023.
76. Maslove DM, Podchiyska T, Lowe HJ. Discretization of continuous features in clinical datasets. 2013;544–553. https://doi.org/10.1136/amiajnl-2012-000929.
77. Li R, Wang Z. An entropy-based discretization method for classification rules with inconsistency checking. First International Conference on Machine Learning and Cybernetics, November, 2002;4–5.
78. World Health Organization [WHO]. Malaria surveillance, monitoring & evaluation: A reference manual. 2018. Geneva-Switzerland.
79. Li G, Zhou X, Liu J, Chen Y, Zhang H, Chen Y, Liu J, Jiang H, Yang J, Nie S. Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Negl Trop Dis. 2018;12(2):1–19. https://doi.org/10.1371/journal.pntd.0006262.
80. Ali MFM, Asklany SA, El-wahab MA, Hassan MA. Data Mining Algorithms for Weather Forecast Phenomena: Comparative Study. International Journal of Computer Science and Network Security. 2019;19(9):76–81.
81. Makhtar M, Nawang H, Shamsuddin SNW. Analysis on Students Performance Using Naïve classifier. J Theor Appl Inf Technol. 2017;95(16), 3993–4000. www.jatit.org.
82. Zhu C, Idemudia C, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. In Informatics in Medicine Unlocked 2019;(pp. 4–5). Elsevier Ltd. https://doi.org/10.1016/j.imu.2019.100179.
83. Simsek S, Kursuncu U, Kibis E, AnisAbdellatif M, Dag A. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. In Expert Systems with Applications 2020;(Vol. 139). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2019.112863.
84. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. In Informatics in Medicine Unlocked. 2018. Elsevier Ltd. https://doi.org/10.1016/j.imu.2017.12.006.
85. Vapnik WN. The nature of statistical learning theory. 2000. Tsinghua University Press.
86. Ahmad L, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A. Informatics using three machine learning techniques for predicting breast cancer recurrence. Health & Medical Informatics. 2013;4(2):2–4. https://doi.org/10.4172/2157-7420.1000124.
87. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: A brief primer. Behavior Therapy. 2020. https://doi.org/10.1016/j.beth.2020.05.002.
88. Titterington M. Neural Networks. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(1):1–8.
89. Wang Q. A Hybrid Sampling SVM Approach to Imbalanced Data Classification. 2014;(Vol. 2014, pp. 1–7). Hindawi Publishing Corporation.
90. Zhao J, Jin J, Chen S, Zhang R, Yu B, Liu Q. Knowledge-Based Systems. Knowl-Based Syst. 2020;203:1. https://doi.org/10.1016/j.knosys.2020.106087.
91. Priya A, Garg S, Tigga NP. Predicting anxiety, depression and stress in modern life using machine learning algorithms machine learning algorithms. International Conference on Computational Intelligence and Data Science. 2019;167:1258–67. https://doi.org/10.1016/j.procs.2020.03.442.
92. Soleymani R, Granger E, Fumera G. F-Measure Curves: A Tool to visualize classifier performance under imbalance. In Pattern Recognition: Elsevier Ltd.; 2019. https://doi.org/10.1016/j.patcog.2019.107146.
93. Patil S, Sonavane S. Improved classification of large imbalanced data sets using rationalized technique : Updated Class Purity Maximization Over _ Sampling Technique (UCPMOT). Journal of Big Data. 2017;4(49):1–32. https://doi.org/10.1186/s40537-017-0108-1.
94. Mehdiyev N, Enke D, Fettke P, Loos P. Evaluating forecasting methods by considering different accuracy measures. Proc Compu Sci. 2016;95:264–71. https://doi.org/10.1016/j.procs.2016.09.332.
95. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. J Eval Clin Pract. 2016;22:835–43. https://doi.org/10.1111/jep.12515.
96. Goetz JN, Brenning A, Petschko H, Leopold P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput Geosci. 2015;81:1–11. https://doi.org/10.1016/j.cageo.2015.04.007.
97. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.
98. Gareth J, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. Springer. 2014.

99. Witten I, Frank E, Hall M. Data mining: Practical machine learning tools and techniques (3rd ed.). 2011. Morgan Kaufmann.
100. R Core Team. R: A language and environment for statistical computing. 2020.https://www.r-project.org/.
101. Casas P. funModeling: Exploratory Data Analysis and Data Preparation Tool-Box (1.9.3). 2019.https://cran.r-project.org/package=funModeling.
102. Wickham H, François R, Henry L, Müller K. dplyr: A grammar of data manipulation (0.8.5). R Foundation for Statistical Computing. 2020. https://cran.r-project.org/package=dplyr.
103. Wickham H, Henry L. tidyr: Tidy Messy Data (1.0.2). R Foundation for Statistical Computing. 2020.
104. Kuhn M. caret: Classification and Regression Training (6.0-86). R Foundation for Statistical Computing. 2020. https://cran.r-project.org/package=caret.
105. Dinov I. Evaluating Model Performance. Data Science and Predictive Analytics. 2020. http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/13_ModelEvaluation.html.
106. Parikh R, Mathai A, Parikh S, Sekhar C, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian Journal of Opthamology. 2008;56(1):45–50.
107. Enke D, Mehdiyev N. A new yybrid approach for forecasting interest rates. Proc Comp Sci. 2012;12:259–64.
108. Ahlawat A, Suri B. Improving Classification in Data mining using Hybrid algorithm. IEEE. 2016;2– 5.
109. Lal A, Kumar CRS. Hybrid Classifier for Increasing Accuracy of Fitness Data Set. International Conference for Convergence in Technology. 2017;1246–1249. https://doi.org/10.1109/I2CT.2017.8226326.
110. Nimala K, ThamizhArasan R. Hybrid data mining approaches for accurate prediction of diabetes and heart disease. International Journal of Pure and Applied Mathematics. 2018;120(6):2693–705.