Journal of Health Informatics in Africa

Reliability of Predictions Using Hybrid Models: The Case of Malaria Incidence Rates in Uganda

Francis Fuller Bbosa[a],[c,*], Ronald Wesonga[b] , Peter Nabende[c], Josephine Nabukenya[c]

[a]School of Statistics and Planning, Makerere University, Kampala, Uganda
[b]Department of Statistics, College of Science, Sultan Qaboos University, Muscat, Oman
[c]School of Computing and Informatics Technology, Makerere University, Kampala, Uganda

Background and purpose: Reliability of estimates emanating from predictive independent data mining techniques is a complex problem. This could be attributed to cross-cutting weaknesses of individual techniques such as collinearity due to high dimensionality of attributes in a dataset, biasedness due to under fitting and over fitting of data as well as noise accumulation due to outliers and thus affecting the reliability of predictions emanating from these models. This study thus aimed at developing a hybrid data mining technique for predicting reliable malaria incidence rate thresholds.
Methods: The decision tree and naïve Bayes classifiers were used to build a hybrid prediction model. Results of the developed hybrid model were compared with independent data mining models using 10-fold cross-validation on a previously unlearned data set. Accuracy, F-measure and the area under the receiver operating characteristics curve (AUC) were the key performance metrics used to evaluate the generalizability of the hybrid model in comparison to the independent models.
Results: Findings revealed that the hybrid classifier attained an accuracy of 79.3% and an F-measure score of 84.2%, the naïve Bayes classifier achieved accuracy and F-measure value of 69% while the decision tree classifier registered an accuracy of 72.4% and an F-measure score of 80%.
Conclusions: The developed hybrid model outperformed both independent decision tree and naïve Bayes models. Hence merging several independent homogeneous predictive data mining techniques enhances the accuracy of the estimates leading to reliable estimates.

Keywords: Hybrid, Data mining, Prediction, Hybrid, Malaria, Incidence

1 Introduction

Garg and Vishwakarma [1] argue that predicting a reliable estimate based on independent data mining techniques is an intricate obstacle, as each technique has its weaknesses with respect to the data structure, shape, and validity [2] [3]. According to Gidron [4], reliability refers to the degree of consistency in measurement. Easterby-Smith, Thorpe, and Jackson [5] corroborate that reliability can be explored by answering the following three questions: i) Do the predicting methods employed generate similar results on different occasions? ii) Are the same results generated by other researchers? iii) Is the analytical process from raw data to the discovery of new knowledge transparent? Hence reliability is a measure of the consistency of the information [6].

Prediction of estimates in databases is often made based on traditional statistical techniques rather than data mining techniques [7] [8] [9] [10] [11] to mine formerly hidden patterns and information from databases. However, the current proliferation of data that is "big" in nature and unstructured, characterized by its Volume, Velocity, Variety, Veracity, and Value have made it difficult for traditional statistical procedures that are often exclusively accustomed to the investigation of structured and homogeneous data, to process and analyze large and complex data sets [12] [13] [14]. The fact that the data is too big and in different forms as well as from various sources led to several scholars [14] [15] [16] breaking down big data into five characteristics, commonly referred to as 5 V's: Volume relates to the size of data, Variety pertains to the data which appears in different forms, Velocity denotes the high pace at which new data is

*Corresponding author address: Department of Planning and Applied Statistics, School of Statistics and Planning, Makerere University, P.o Box 7062, Kampala, Uganda. Email: fullerbbosa@gmail.com. Tel: (+256)-702792202
 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

generated, Veracity measures the authenticity of the data, and Value assesses how good the quality of the data is in reference to the intended results. Therefore, the rise of big data has forced scholars [13] [17] [18] to advance data mining as a plausible solution to extract previously unknown and unseen patterns and information that are challenging to discover with traditional statistical techniques with respect to big data.

Agyapong, Hayfron-Acquah, and Asante [19] assert that data mining techniques are mainly categorized into two categories: predictive and descriptive methods. Predictive approaches also known as classification learn 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 [20] [21] whereas descriptive approaches are also known as clustering usually identify patterns or associations among attributes in datasets by looking for human-interpretable patterns that describe data [19].

Pertinent literature reveals that predictive approaches are the dominant method in the data mining arena [20] [22] [23] possibly due to their strength such as making the computation process easy to understand, generation of inclusive rules for classification, handling both real and discrete data [21] [22] [24]. However, majority of the independent predictive techniques share common weaknesses such as dependence on the nature of the dataset or data type for classifier performance [22], imprecision of estimates in scenarios where various attributes in a dataset are dependent on each other [21] , replication of sub-trees on different paths leading to collinearity [25], information overload due to the large size of input datasets thus increasing the time to mine information, which decelerates the decision-making process [25], collinearity due to high dimensionality of attributes in a dataset [26], biasedness due to under fitting and over fitting of data as well as noise due to outliers [27]. Thus several researchers [26] [27] [28] suggest that different independent data mining models have varying predicting capabilities based on their strengths and weakness; the authors claim there is no universally employable independent data mining model for all prediction scenarios. Hence the practice of employing independent data mining techniques leads to unwanted biases, errors and omissions, noise accumulation and spurious correlations among variables which affects the accuracy and reliability of predictions emanating from these models [29] [30] [31].

Various scholars [32] [33] [34] have applied more than one independent data mining technique to predict estimates on the same dataset but all these techniques generated varying results with dissimilar accuracies. As a result, the above scholars conclude that there is no single data mining model that produces the most reliable result. To address the above gap associated with variances in estimates of predictions using individual classifiers, hybridization of several individual data mining techniques is suggested [25] [35] [36] [37] alluding to the fact that merging several independent data mining techniques improves the accuracy of the estimates leading to reliable estimates [25] [28] [34] [38]. According to Ahlawat and Suri [25], hybrid procedures in data mining are a logical amalgamation of various individual techniques, thereby utilizing the strengths of the individual procedures of the hybrid algorithm to improve the performance of prediction models to generate reliable estimates. Kazienko, Lughofer, and Trawiński [39] suggest that it's imperative to note that both hybrid and ensemble techniques utilize the concept of information amalgamation nonetheless in diverse ways. In case of hybrid classifiers, diverse heterogeneous data mining approaches are combined [39] [40] [41] whereas ensemble classifiers instead merge numerous but homogeneous, feeble techniques [42], characteristically at their individual output level, utilizing several merging methods [43].

1.1 Data mining models employed in the study

Despite the presence of several predictive data mining techniques, scholars are facing the challenge of choosing the best model for a particular data set [44]. In keeping with relevant published literature, the most frequently employed predictive data mining techniques include: Decision trees, Artificial Neural Networks (ANN), KNearest Neighbor (k-NN), Support Vector Machines (SVM), algorithm, logic-based algorithms especially Decision Trees (DT) and bayesian related classifiers [45]. Furthermore, Hamblin et al. [45] reveal that ANN and SVM generate better estimates when dealing with continuous-valued attributes whereas K-NN is biased to noise and hence very sensitive to outliers in datasets. However, given that the researchers' problem under investigation involved discrete data from heterogeneous sources, the above limitations disqualify ANN, SVM, and KNN techniques.

However, logic-based systems such as Bayes and decision trees classifiers tend to perform better when dealing with categorical attributes [45]. As a result, the researchers employed the decision tree and naive Bayes as the predictive data mining techniques for this study on the basis of their greater ability of

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

modelling classification type prediction problems [46]. Above all, the choice of decision trees and Bayesian classifiers takes into consideration decision boundary-based and probability-based approaches to prediction in machine learning respectively [47].

1.2 Malaria disease model

Malaria was chosen as a disease model for this study because the World Health Organization (WHO) [48] recognizes the presence of weak surveillance systems that are unable to reliably predict future malaria incidence rates particularly in Low and Middle-Income Countries (LMICs) particularly Uganda, making it hard to optimize response to malaria outbreaks. Additionally, the latest WHO world malaria report [49] reveals that at a global level, Africa accounted for 213 million cases out of the 228 million cases recorded globally in 2018 with Uganda accounting for 5% of the global burden. Despite increased consideration paid to malaria surveillance systems and their key role for improving health systems in Low and Middle-Income Countries (LMICs) such as Uganda, it is believed that the majority of the existing surveillance systems cannot be used to reliably predict future malaria incidence rates [49]. Hence the need to develop a robust hybrid prediction model in order to enhance early warning leading to effective and timely response to future outbreaks.

The overall motivation underlying this study is that considerable work has been done on boosting the predictive accuracy of individual homogeneous data mining techniques but little work with regards to enhancing the reliability of heterogeneous techniques. To this end, the researchers argue that there is a need for building a hybrid data mining approach, which is an effective amalgamation of numerous independent data mining techniques, to utilize the strengths of each individual technique and compensate for each other's weaknesses.

1.3 Problem statement

In order to address the drawbacks of the traditional statistical methods, data mining techniques have been adopted. However, the conventional independent data mining techniques are not capable of producing reliable predictions. This is mainly attributed to their weaknesses with respect to the data structure, shape, and validity. As a result, the performance of conventional independent data mining techniques is weakened due to noise accumulation emanating from measurement errors, outliers, and missing values as well as spurious correlations which may lead to false scientific conclusions or poor predictions and varying measures of accuracy. Hence, the need to develop a hybrid data mining algorithm in order to improve the predictive accuracy and reliability of individual data mining techniques.

2 Related Literature

In this section, the researchers review recent research on the amalgamation of independent data mining for various real-world predictive problems.

Sumana and Santhanam [44] examined single and hybrid classification techniques as viable tools to achieve enhanced predictions for the presence or absence of heart diseases in a cohort of patients. Their findings reveal that the proposed hybrid model produced more accurate estimates of 99.54% compared to single classifiers and ensemble classifiers.

Ogwoka, Cheruiyot, and Okeyo [50] proposed "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms". Findings reveal that merging Decision tree and k-means algorithms generated better results after extracting previously unknown features, thus improving the accuracy of prediction.

In 2016, Dubey and Saxena [51] developed a hybrid prediction model for feature selection. They amalgamated correlation and support vector machines to classify big data. This proposed hybrid technique was tested on five big data boolean datasets. The authors attest that the hybrid yielded better accuracies in three out of the five big data datasets with a fewer number of features.

Ahlawat and Suri [25] developed a hybrid algorithm by combining decision trees and clustering to classify data samples. Their results proved that an amalgamation of decision trees and clustering is suitable

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

to improve the accuracy of estimates. They concluded that using hybridization can be used to enhance performance and prediction values to get better results.

In 2016, Raghavendra and Indiramma [52] proposed a "Hybrid data mining model for the classification and prediction of medical datasets". The researchers employed attribute separation selection techniques particularly the forward selection and backward elimination method generate an appropriate subset of attributes to enhance the performance of the model. Findings revealed that the proposed hybrid model outperformed the linear regression and artificial neural networks with fewer number of significant attributes.

Hakizimana et al. [2] proposed "A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets". The authors built a hybrid model for predicting infections occurrence in datasets by merging naïve Bayes, random forest, simple logistic, Bayesian logistic regression, and SMO. The hybrid technique produced the best accuracy with 100%, compared to the naïve Bayes with 90.9%, SMO with 90.9%, and Bayesian logistic regression with 36.4%. Hence the hybrid model is superior to individual models in terms of improved accuracy of estimates.

Ren, Fei, Liang, Ji, and Cheng [53] in their study to predict kidney disease in hypertension patients proposed a hybrid neural network that integrates Bidirectional Long Short-Term Memory (BiLSTM) and Autoencoder networks. Findings from their study revealed that the proposed hybrid model attains 89.7% accuracy and thus the proposed integrated model outperformed traditional stand-alone prediction models with distinct features and neural baseline systems.

In order to predict diseases, [54] developed a hybrid model that amalgamated k-nearest neighbor,case-based reasoning, and fuzzy set classifiers. Findings revealed that the hybrid model enhanced the accuracy of the model compared to the stand-alone classifiers. The authors concluded that the integration of several independent predictive models aids to yield improved estimates from predictions. In 2020, Ju-young, Rang, and Jong-chul [55] proposed "Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management". The researchers used a hybrid model that integrated multiple regression and artificial neural networks to forecast mean daily temperature. The hybrid model yielded results with a root mean square error (RMSE) of 1.02-3.35 compared to the standard climate model that achieved a RMSE of 1.61-3.37.

Junliang Fan, Wu, Ma, Zhou, and Zhang [56] suggested three novel hybrid support vector machines (SVM) with bat algorithm (SVM-BAT), whale optimization algorithm (SVM-WOA) and particle swarm optimization algorithm (SVM-PSO) for daily diffuse solar radiation in air-polluted regions using, multivariate adaptive regression spline (MARS), SVM and extreme gradient boosting (XGBoost) models. Their findings showed that hybrid models generate more accurate estimates. The researchers proved that the hybrid SVM-BAT model was a better classifier than the SVM, XGBoost, and MARS models by attaining more accurate daily rates and quicker convergence rates.

A synopsis of the reviewed relevant literature suggests that the application of various independent predictive data mining techniques in the context of classification on similar datasets yields varying estimates, leading to poor, unreliable estimates and consequently insufficient scientific conclusions. The above shortcomings have stimulated the curiosity of the researchers to continuously struggle to improve the algorithms for undertaking classifications and predictions using single data mining techniques related to the realization of reliable estimates. To this end, the researchers argue that there is a need for building a hybrid data mining approach, which is an effective amalgamation of numerous independent data mining techniques, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses.

3 Materials and methods

3.1 Data Pre-processing

3.1.1 Data Collection .

Monthly data for the period January 2012 to December 2019 on confirmed and suspected (clinically diagnosed) cases of malaria were obtained by the researchers from the Ministry of Health through the

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

District Health Information System 2 (DHIS2)[1]. Temperature (average maximum and average minimum) and rainfall data for a similar period were also obtained from the Uganda National Meteorological Authority (UNMA)[2], whereas demographic data was obtained from the Uganda Bureau of Statistics (UBoS)[3].

3.1.2 Data Cleaning

The researchers verified and validated the raw datasets in order to check for errors, omissions, and outliers in preparation for compiling a complete and merged dataset that was used for building predictive models. R version 3.6.3 (R Core Team, 2020) served as the primary tool for data management. In cases of missing climate data, the researchers imputed the missing data by substituting each missing value with the average of identified values of that attribute using equation (1) adapted from [57];

$$Y_i^j = \sum k \in r(\text{complete}) \frac{Y_k^j}{n_{|r(\text{complete})|}} \quad (1)$$

Where Y_i^j denotes the j^{th} the missing attribute of the i^{th} observation
$r(complete)$ denotes non-missing values from Y_i
$r_{|I(complete)|}$ denotes the total number of observations where the j^{th} attribute is not missing.

3.1.3 Data Transformation

Normalization.
The fact that attribute values were measured on different scales .e.g. temperature in degrees Celsius and rainfall in milimetres implied that the attributes couldn't be compared meaningfully [58]. Hence data normalization by standardization (z-scores) was undertaken to adjust for the above discrepancies, thereby ensuring that all continuous attribute values are scaled and belong in similar ranges [58] [59]. The mean and standard deviation of the attributes were used for normalization as illustrated in equation (2);

$$B = \frac{x_i - \mu}{\sigma} \quad (2)$$

where $B = normalised\ attribute\ value, x_i = original\ attribute\ value$
$\mu = mean\ of\ attribute$, $\sigma = attribute\ standard\ deviation$

Discretization.
According to [60], data discretization enhances the comprehensibility of the discovered previously unknown knowledge from databases. Hence data discretization was undertaken in order to produce a homogeneous group of antecedent attributes since the dataset comprised both continuous and categorical attributes, thus alleviating outliers and conquering noise accumulation [59] [60] [61].

3.1.4 Training and Testing data

The dataset was split into training and testing datasets. 80% of the data was assigned to the training group for the development of the classifiers. The rest of the data (20% of the total cases) was assigned to the validation groups for the assessment of model performance [62].

3.2 Ethical statement on data access

The datasets were accessed with official permission granted from the Ministry of Health, Uganda National Meteorological Authority and Uganda Bureau of Statistics.

[1] www.health.go.ug
[2] www.unma.go.ug
[3] www.ubos.org

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

3.3 Analysis

3.3.1 Proposed hybrid data mining model

The researchers' proposed hybrid model was developed in two phases, utilizing two different single techniques. In the first phase, decision trees were used in a cascaded style for important attribute extraction based on the gain ratio to pre-process the data. Then, the output of the first stage was employed to construct the second stage weighted naïve Bayesian classifier as the prediction model. The overall methodological workflow of the hybrid model is illustrated in Figure 1.

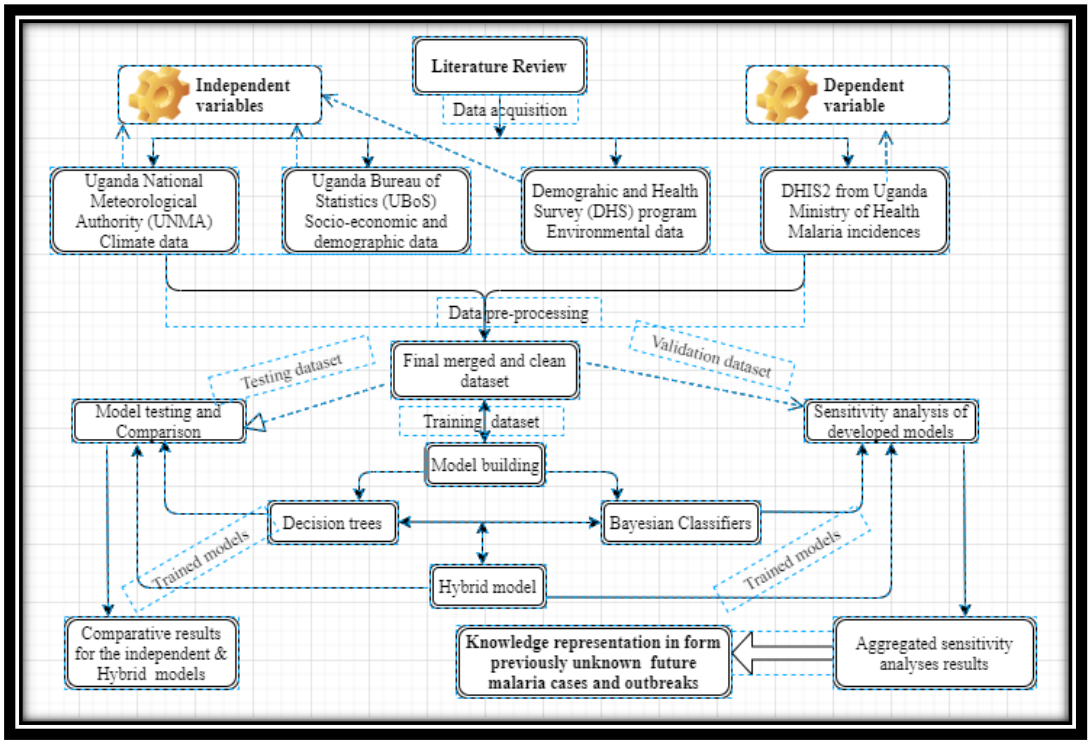

Figure 2: The overall framework and methodological workflow of the hybrid model

First Phase

Under this phase, the researcher employed the decision tree technique based on the C4.5 algorithm to identify the most significant attributes to improve estimates [63]. Additionally, the researchers assigned gain ratio values as weights for each attribute based on the fact that weighted classification assigns various degrees of significance to different attributes and classes in order to denote the relative importance of each attribute and class [64].

The following procedures were followed in the first phase.

i) Given a training dataset, T with instances y_i where $y_i = \{y_1, y_2, \dots, y_n\}$.The training dataset T is defined by attributes B_i .i.e. $T = \{B_1, B_2, \dots, B_n\}$ and generate an attributes list B which has v possible values. The training data also belongs to a set of classes $Z = \{Z_1, Z_2, \dots, Z_n\}$

ii) Build a decision tree classifier employing the C4.5 formula adapted from [65];

a) Entropy for the root node:

$$\text{Entropy}(T) = \sum_{i=1}^{m}(p_i)\log_2(p_i)$$

Where p_i denotes the probability of the target attribute

b) Entropy of Attribute *(B)* for attribute list B with respect to the root note *(T)*

$$\text{Entropy}_B(T) = \sum_{j=1}^{v}\frac{|T_j|}{|T|} * \text{Entropy}(T_j)$$

where T_j is a collection of the instances in the dataset T with attribute B having value j

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

c) Compute the information gain for each attribute

$$\text{Info\_Gain(T)} = \text{Entropy(T)} - \text{Entropy}_{\text{B}}\text{(T)}$$

d) Compute the split information *(V) for a set of attributes* (T_i) *and* (T_j)

$$\text{Splitinfo(B)} = -\left[\left|\frac{\text{T}_\text{i}}{\text{B}}\right| \log_2 \left[\left|\frac{\text{T}_\text{i}}{\text{B}}\right|\right] + \left|\frac{\text{T}_\text{j}}{\text{B}}\right| \log_2 \left[\left|\frac{\text{T}_\text{j}}{\text{B}}\right|\right]\right]$$

e) Compute the gain ratio*(B)* for attribute list *B w*ith respect to the root note *(T)*

$$\text{Gainratio(B)} = \left(\frac{\text{Info\_Gain(B)}}{\text{Splitinfo(B)}}\right) \quad (3)$$

f) After computing information gain ratio for each attribute on the decision tree classifier, assign and initialize the weight (W_i) for each attribute(B_i), where $B_i \in T$ as the gain ratio value of the respective attribute.

NB: The weight for an attribute is computed as $W_i = Gainratio(B)_i$

g) If the attribute, $B_i \in T$, is not tested in the decision tree, then the weight (W_i) of the attribute, (B_i), was initialized to zero.

h) Thus, the parent node of the tree will have a higher weight value in comparison with those of its child nodes [66].

Second Phase

Under this phase, the researchers employed the naïve Bayes technique. The naïve Bayes technique is a Bayesian classifier grounded on statistical methods and utilizes Bayes Theorem proposed by Thomas Bayes to calculate unknown conditional probabilities [67]. Bayesian classifiers handle real and discrete data and make the computation process easier. The main advantages of naive Bayes classifiers are that they are resilient to noise and outliers, and they handle missing values by ignoring the instance during probability estimate calculations [68]. The naïve Bayes technique is referred to as "naïve" due to the fact that it assumes that the occurrence of a certain attribute is independent of other attributes conditional on a similar consequent target value [67].

On the other hand, the naïve Bayes technique assumes conditional independence of antecedent attributes given the target attribute [69] [70], which hardly ever holds in real-world scenarios [71]; weakening its performance in models with complex attribute dependencies [72].

Hence in order to alleviate the independence hypothesis of the naive Bayes, the researchers applied weight values derived from the first phase to the attribute set based on each attribute's importance in the classification process [45]. According to [45], the Naive Bayes algorithm is derived from Bayes' theorem (equation 4);

$$\text{P(X/Y)} = \frac{\text{P(Y/X)P(X)}}{\text{P(Y)}} \quad (4)$$

Where P(X/Y)=probability of X given Y has occurred

P(Y/X)=probability of Y given X has occurred

P(X) and P(Y) are probabilities of X and Y occurring independently from each other.

However, bases on the assumption of independence among antecedent attributes, equation (4) is transformed to equation (5) for the naïve Bayes formula [67];

$$\text{P}(\text{Z}_\text{n}/\text{y}_1, \text{y}_2, \ldots, \text{y}_\text{n}) = \frac{\text{P}(\text{y}_1, \text{y}_2.\text{y}_3, \ldots., \text{y}_\text{n}, \text{Z}_\text{i})}{\prod_{\text{i}=1}^{\text{n}} \text{P}(\text{y}_\text{i})} \quad (5)$$

Where $\text{P}(Z_n)$= the prior probability of the class that reflects background knowledge due to the chance of Z to be correct.

$\text{P}(y_i)$= the probability of y to be observed

$\text{P}(\text{Z}_\text{n}/\text{y}_\text{i})$= the posterior probability of class (malaria incidence) given predictor (attribute).

$\text{P}(\text{y}_\text{i}/\text{Z}_\text{n})$= the probability of observing y given Z holds

Hence simplifying the numerator on the right hand side in equation (5) leads to equation (6);

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

$$P(Z_n/y_1, y_2, \dots, y_n) = \frac{P(y_n)\prod_{i=1}^{n} P(y_i/Z_n)}{\prod_{i=1}^{n} P(y_i)} \quad (6)$$

Since the denominator in equation (6) is invariant across various consequent attribute classes, it can be dropped as illustrated in equation (7).

$$x = \text{argmax}(P(Z_n/y_i) = P(y_n)\prod_{i=1}^{n}\prod P(y_i/Z_n)) \quad (7)$$

Where x is the class with the highest probability given a set of attributes.
The following procedures were followed in the second phase.

i) The researcher will compute the class conditional probabilities utilizing only the significant attributes nominated by decision tree technique in the first phase (i.e. $W_i \neq 0$) and classify each instance $B_i \in T$ based on the gain ratio values.

ii) Assume that there are m classes, $Z_1, Z_2, \dots, Z_m$. Given an object Y, the classifier will predict that B belongs to the class having the highest posterior probability.

That is, the naïve Bayesian classifier predicts that tuple B belongs to the class Z_i if and only if

$$P(Z_i/Y) > P(Z_j/Y) \text{ for } 1 \leq j \leq m, j \neq i$$

iii) Thus the $P(Z_i/Y)$ needs to be maximized. The class Z_i for which $P(Z_i/Y)$ is maximized is called the maximum posterior hypothesis.

iv) Compute the class conditional probabilities using the weights of significant attributes as exponential constraints using the formula below [72];

$$P(Z/y_i) = P(Z)\prod_{j=1}^{n}\prod P(y_i|Z)^{Wi} \quad (8)$$

Where W_i refers to the weight of the attribute, (y_i), which effects on class conditional probability calculation as an exponential parameter.

$P(y/Z) = Probability\ of\ attribute\ y\ given\ Z\ has\ occured$
$P(Z) = Probability\ of\ \ consequent\ attribute(class)$
$P(y) = Probability\ of\ \ antecedent\ attribute$
$P(Z/y) = Probability\ of\ event\ Z\ \ given\ y\ has\ occured$

v) The class conditional probabilities of the non-significant attributes ($W_i = 0$) by decision trees will not be employed in the prediction of estimates in the second phase.

vi) The researcher reiterated this process until all the attributes were correctly predicted.

The algorithm in Figure 2 adapted from [63] [73] outlines the proposed hybrid algorithm;

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

Input: Training dataset $T = \{y_1, y_2, \ldots, y_n\}$
Output: Hybrid model
1: Determine the best splitting attribute;
2: Create a root node $\{Z\}$;
3: Z=Generate root node **arc** for each split base;
4: **for** $arc \in Z$ **do**
5: D =dataset generated by employing splitting base to D;
6: **if** stopping basis achieved for this path,
7: Create a leaf node (T);
8: **else**
9: T=rebuild D;
10: **end if**
11: Z =Add T to arc;
12: **end for**
13: $W = \{w_1, w, \ldots, w_n\}$//weights for $y_i \in T$;
14: **for** $y_i \in T$ **do**
15: **if** y_i is not tested in T ;
16: $(W_i = 0)$;
17: **end if**
18: **end for**
19: **for** $Z_i \in T$ **do**
20: Determine prior probabilities $P(Z_i)$;
21: **end for**
22: **for** $B_i \in T$ and $(W_i \neq 0)$**do**
23: Determine the conditional probabilities $\boldsymbol{P(B_{ij}/Z_i)^{W_i}}$
24: Hence compute posterior probability $P(Z_i/y_i)$
25: **end for**

Figure 2: Proposed hybrid algorithm

3.4 Goodness of Fit

The researchers used k-fold cross-validation (CV) method and six performance evaluation metrics.

3.4.1 Classifier Validation Method

The K-fold cross-validation method was employed [74]. In k- fold validation, the set of training data was divided into k- groups of equal size. In our experiment, we used the K=10 cross-validation due to the fact that its performance is reliable [60]. Hence under the 10-fold cross-validation process, 90% of the data was used for training and 10% of the data was used for testing purposes.

3.4.2 Performance Evaluation Metrics

The researchers used a confusion matrix in order to evaluate the performance of the classifiers based on various performance evaluation metrics as illustrated in Table 1.

Table 5: Performance metrics computed

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

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

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
Area under the Curve (AUC)		The area under the curve (AUC) is a model goodness-of-fit measure that compares it to a baseline 50% measure (the straight line).

Source: Mehdiyev, Enke, Fettke, & Loos [75]

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".
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".

3.5 Software Tools

The data processing and analysis was undertaken entirely in R, version 3.6.3 [76], by means of R packages "funModeling" version 1.9.3 [77], "dplyr" version 0.8.5 [78], "tidyr" version 1.0.2 [79], "caret" version 6.0.86 [80], "reshape2" version 1.4.4 [81].

4 Results

The overall performance of the classifiers was evaluated based on their prediction accuracy in classifying the instances of the data set into low and moderate malaria incidence thresholds. The researchers utilized 10-fold cross-validation to assess the performance of the three classifiers on previously unlearned data. Figure 3 shows the classification results of the test data.

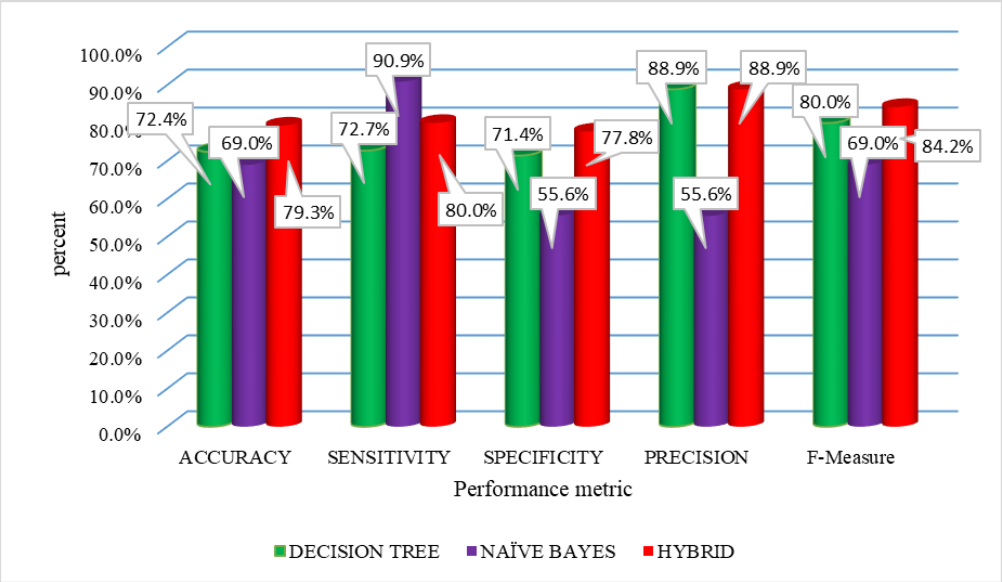

Figure 3: Comparison of classifiers' performance using 10 fold cross-validation

The performance metrics for all the classifiers were separately identified using trained models that were fit on previously untrained test data. According to figure 2, the hybrid model attained the highest performance with respect to the accuracy, specificity, and F-measure metrics recorded at 79.3%, 77.8%,

and 84.2% respectively. On the other hand, the naïve Bayes classifier registered the highest sensitivity at 90.9% compared to 72.7% registered by decision tree and 80% obtained by the hybrid classifier. The achieved accuracy results indicate that the proposed hybrid model outperformed the independent decision tree and naïve Bayes classifiers by 6.9% and 10.3% respectively.

4.1 Receiver Operating Characteristics (ROC) Curve

The ROC curve (Figure 4) is a graphical plot that symbolizes how the performance of the sensitivity and specificity of a classifier varies in relation to one another (Wu, Yang, Huang, He, & Wang, 2018; Zhu, Idemudia, & Feng, 2019). The ROC permitted the researchers to assess the performance of the developed models at various thresholds. Figure 3.0 reveals that the Area under the Curve (AUC) was recorded at 67.17%, 88.38%, and 86.87% for the decision tree, naïve Bayes, and hybrid classifier respectively. A random model would have an AUC of 50% (the straight line), give that it basically dissects the graph (Winters, 2015). Hence the generated the ROC curve for all the classifiers outperform a random model (straight line).

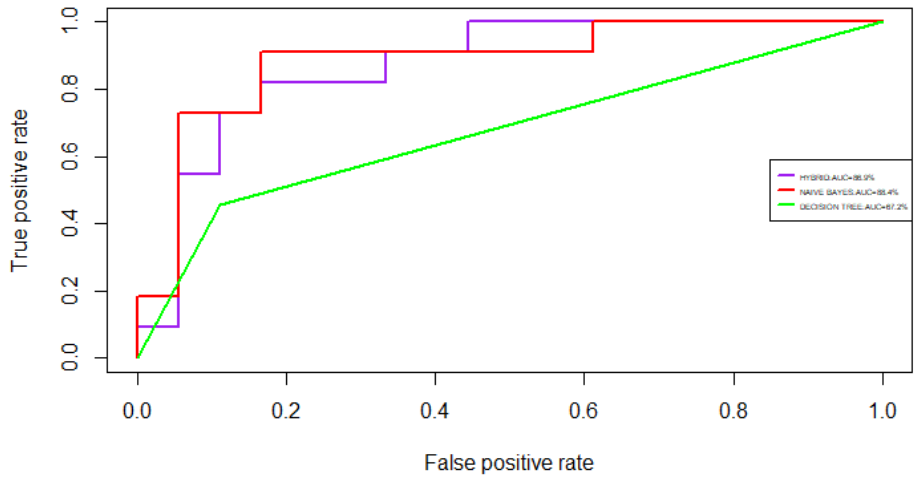

Figure 4: Comparison of the ROCs for the classifiers at various thresholds

Based on Figure 4, the hybrid and naïve Bayes classifiers attained a high sensitivity (True positive rate) of approx.70% at a very low false-positive rate (1-specificity). Nevertheless, the hybrid model denoted by the purple line returned a better cut-off determination threshold since it yielded a higher true positive rate at lower false-positive rates compared to the naïve Bayes classifier. The decision tree was a poor classifier given that it achieved a high True positive rate at the cost of a high false-positive rate.

4.2 Reliability of the proposed hybrid model

To further demonstrate and evaluate the reliability of the developed hybrid model on high dimensional data, the researchers applied the model on three demonstration datasets from various application domains. The datasets were sourced from the UCI machine learning repository[4] [82]; an assembly of databases used by several scholars [83] [84] [85] [86] for experimental investigation of machine learning techniques. Table 2 shows the characteristics of the demonstration datasets.

[4] https://archive.ics.uci.edu/ml/datasets.php

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

Table 2: Datasets used to test the reliability of the hybrid model

Dataset name	Description	Number of attributes	Number of observations	Target attribute	Data Source
Heart failure	Dataset for predicting mortality caused by Heart Failure	13	299	Death_Event	https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
Heart_UCI	Data set with attributes to detect the presence of heart disease in the patient	14	303	target	https://archive.ics.uci.edu/ml/datasets/Heart+Disease
Wine quality	Data set with attributes to determine which physiochemical properties make a wine 'good'!	12	1599	Quality (=<6.5="BAD" >6.5="GOOD"	https://archive.ics.uci.edu/ml/datasets/wine+quality

The researchers subjected the datasets in table 2 to data pre-processing steps similar to the malaria incidence rate dataset collected from a known population in Kampala. The researchers employed k-fold cross-validation to compare the performance of the developed hybrid model in terms of the F-measure metric with the independent decision tree and naïve Bayes classifiers. The results are shown in Table 3.

Table 3: Performance of the decision tree, naive Bayes and proposed hybrid models on various demonstration datasets

Dataset	Model	Accuracy	Sensitivity	Specificity	Precision	F-measure
heart_failure	Decision tree	81.7%	86.4%	68.8%	88.4%	87.4%
	Naïve Bayes	81.7%	90%	65%	83.7%	86.7%
	Hybrid	86.7%	90.7%	76.5%	90.7%	90.7%
heart	Decision tree	83.6%	89.5%	81%	68%	77.3%
	Naïve Bayes	83.6%	75.9%	90.6%	88%	81.5%
	Hybrid	88.5%	82.1%	93.9%	92%	86.8%
winequality	Decision tree	84.1%	86.8%	52%	95.5%	90.9%
	Naïve Bayes	79.7%	88.9%	39%	86.6%	87.7%
	Hybrid	84.7%	88.2%	54.5%	94.4%	91.2%

Table 3 reveals that the hybrid model improved the performance the independent models across all the datasets. The proposed hybrid model outperformed the independent models by obtaining the highest F-measure score of 90.7%, 86.8% and 91.2% on the "heart_failure", "heart" and "winequality" datasets respectively. Similarly, the proposed hybrid model outperformed the independent models in terms of predictive accuracy in all the demonstration datasets. The attained results indicate that the proposed hybrid model could help in enhancing the performance of the independent data mining techniques.

5 Discussion

The main purpose of this study is to build a hybrid data mining approach robust to noises, dependence, and data complexity to improve the predicting of malaria incidence rates, leading to early prediction of malaria occurrences and thus dipping the transmission risk in the community. In this work, the researchers

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289

compared a hybrid data mining technique with single data mining techniques in the form of decision trees and naïve Bayes classifiers. The results showed that the hybrid model outperformed the independent models in terms of classification accuracy, specificity, and F-measure. To assess the robustness of the hybrid model, the researchers undertook further experiments using datasets from different application domains. These datasets were obtained from the UCI machine learning repository [82]. Findings from these experimental analyses alluded to the researchers' initial findings with the hybrid model outperforming the individual models. Furthermore, the experimental results showed that employing the hybrid model by weighting the naïve Bayes classifier using gain ratio values emanating from the C4.5 algorithm enhances the naïve Bayes algorithm. This is similar to findings of [72] [87] [88] [89] who alluded that nullifying the conditional independence assumption of the naïve Bayes through weighting can help improve its performance.
Above all, the results are in agreement with the concept that high sensitivity and specificity may not be achievable in real-world scenarios concurrently [90] because they are inversely related, implying that as the specificity increases, the sensitivity decreases and vice versa [91]. Hence there is a trade-off between sensitivity and specificity with the hybrid model recording a lower sensitivity of 80% compared to the naive Bayes model registered at 90.9%. However, a similar trend was observed in terms of specificity with the hybrid model registering the highest specificity of 77.8% compared to 71.4% and 55.6% recorded for the decision tree and naive Bayes models respectively.

The study faced a key challenge of available monthly data being limited to a few predictor attributes for the period under investigation and hence the researchers were unable to subject the developed hybrid model to a higher dimensional dataset from a known population which would return more reliable and robust performance results [92]. Additionally, the researchers did not take into consideration the effect of biasedness associated with predictions emanating from imbalanced data [93].

6 Conclusion

This study aimed to develop a hybrid model for predicting reliable estimates of malaria incidence thresholds. After reviewing relevant literature, the researchers proposed a hybrid model which was an amalgamation of the C4.5 decision tree and naïve Bayes classifiers. The hybrid model was developed in two phases with phase one employing the C4.5 algorithm to generated information gain ratio values for antecedent attributes which were used as attribute weights for the naïve Bayes classier in the second phase. Empirically, the developed hybrid model outperformed both independent decision tree and naïve Bayes models. Notably, the hybrid model outperformed the independent decision tree and naïve Bayes classifiers in terms of accuracy by 6.9% and 10.3% respectively. Hence merging several independent homogeneous predictive data mining techniques enhances the accuracy of the estimates leading to reliable estimates.

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).

Conflict of Interest

The authors declare that they have no competing interests.

References

[1] Garg, P., & Vishwakarma, S. K. (2019). An efficient prediction of share price using data mining techniques. International Journal of Engineering and Advanced Technology, 8(6), 3110–3115. https://doi.org/10.35940/ijeat.F9085.088619

[2] Hakizimana, L., Cheruiyot, K., Kimani, S., & Nyararai, M. (2017). A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets. International Journal of Computer (IJC), 27(1), 69–83.
[3] Tan, J., & Wang, F. (2017). A Hybrid Mining Approach to Facilitate Health Insurance Decision : Case Study of Non-Traditional Data Mining Applications in Taiwan NHI Databases. Proceedings of the 50th Hawaii International Conference on System Sciences, 3253–3262.
[4] Gidron, Y. (2013). Reliability and Validity. In M. D. Gellman & J. R. Turner (Eds.), Encyclopedia of Behavioral Medicine. Springer Science+Business Media. https://doi.org/10.1007/978-1-4419-1005-9
[5] Easterby-Smith, M., Thorpe, R., & Jackson, P. R. (2008). Management Research (3rd ed.). Sage.
[6] Curry, A., Flett, P., & Hollingsworth, I. (2006). Managing information and systems:The Business Perspective. Routledge, New York-London.
[7] Appiah, S. ., Otoo, H., & Nabubie, B. (2015). Times series analysis of malaria cases in Ejisu- Juaben Municipality. International Journal of Scientific and Technology Research, 4(06).
[8] Carillo, M., Largo, F., & Ceballos, R. (2018). Principal Component Analysis on the Philippine Health Data. International Journal of Ecological Economics and Statistics, 39(1), 91–96.
[9] Hussien, H. H. (2019). Malaria's association with climatic variables and an epidemic early warning system using historical data from Gezira State , Sudan. Heliyon, 5. https://doi.org/10.1016/j.heliyon.2019.e01375
[10] Muwanika, F., Atuhaire, L., & Ocaya, B. (2017). Journal of Medical Diagnostic Prediction of Monthly Malaria Incidence in Uganda and its Implications for Preventive Interventions. Journal of Medical Diagnostic Methods, 6(2). https://doi.org/10.4172/2168-9784.1000248
[11] Twumasi-Ankrah, S., Wa, P., Nyantakyi, K., & Addo, D. (2019). Comparison of Statistical Techniques for Forecasting Malaria Cases in Ghana. Journal of Biostatistics and Biometric Applications, 4(1), 1–9.
[12] Basco, A., & Senthilkumar, N. C. (2017). Real-time analysis of healthcare using big data analytics. IOP Conference Series: Materials Science and Engineering, 263(4). https://doi.org/10.1088/1757-899X/263/4/042056
[13] Hu, C. H., Lee, H. S., Lara, E., & Gan, S. (2018). The Ensemble and Model Comparison Approaches for Big Data Analytics in Social Sciences. Practical Assessment, Research & Evaluation, 23(17).
[14] Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., & Anthony, L. (2015). Big data in global health : improving health in low- and middle-income countries. Big Data in Health Care, January, 203–208.
[15] Bains, J. K. (2016). Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges. International Journal of Advanced Research in Computer Science and Software Engineering, 6(4). www.ijarcsse.com
[16] Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1(2). http://www.journalofbigdata.com/content/1/1/2
[17] Aparna, K., Reddy, C. S., Prabha, S., & Srinivas, V. (2014). Disease prediction in data mining techniques. International Journal of Computer Science and Technology, 5(2), 17–21.
[18] Sahay, S. (2016). Big data and public health: Challenges and opportunities for low and middle income countries. In Communications of the Association for Information Systems (Vol. 39). https://doi.org/10.17705/1CAIS.03920
[19] Agyapong, K. B., Hayfron-Acquah, J., & Asante, M. (2016). An Overview of Data Mining Models (Descriptive and Predicitve). International Journal of Software & Hardware Research in Engineering, 4(5), 53–60. https://doi.org/10.1007/978-3-319-13084-2_59
[20] Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications, 6(2).
[21] Krishnaiah, V., Narsimha, G., & Subhash, C. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies, 4(1), 39–45.
[22] Gorade, S., Ankit, D., & Preetesh, P. (2017). A Study Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology, 4(1), 210–215. https://doi.org/10.21884/ijmter.2017.4031.zt9tv
[23] Thorat, S., & Kute, S. (2014). Medical Data Mining Life Cycle and its Role in Medical Domain. International Journal of Computer Science and Information Technologies, 5(4), 5751–5755
[24] Ying-ying, W., Yi-bin*, L., & Xue-wen, R. (2017). Improvement of ID3 Algorithm Based on Simplified Inform ation Entropy and Coordination Degree. IEEE, 1526–1530.

[25] Ahlawat, A., & Suri, B. (2016). Improving Classification in Data mining using Hybrid algorithm. IEEE, 2–5.
[26] Kumar, P., & Wahid, A. (2015). Performance Evaluation of Data Mining Techniques for Predicting Software Reliability. International Journal of Computer and Systems Engineering, 9(8), 1946–1953
[27] Anwar, H., Qamar, U., & Qureshi, A. W. (2014). Global optimization ensemble model for classification methods. Scientific World Journal, Vol. 2014. https://doi.org/10.1155/2014/313164
[28] Abuassba, A. O. M., Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines. Computational Intelligence and Neuroscience, Vol. 3405463. https://doi.org/10.1155/2017/3405463
[29] Fan, Jianqing, Han, F., & Liu, H. (2014). Challenges of Big Data analysis. In National Science Review (Vol. 1). https://doi.org/10.1093/nsr/nwt032
[30] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[31] La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. The Condor, 120(2), 414–426. https://doi.org/10.1650/condor-17-206.1
[32] Almarabeh, H., & Amer, E. (2017). A Study of Data Mining Techniques Accuracy for Healthcare. International Journal of Computer Applications, 168(3), 12–17. https://doi.org/10.5120/ijca2017914338
[33] DEEPAK, S. (2016). Knowledge Discovery With Hybrid Data Mining Approach. DAYALBAGH EDUCATIONAL INSTITUTE.
[34] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, Vol. 19, pp. 1–16. https://doi.org/10.1186/s12911-019-1004-8
[35] Lal, A., & Kumar, C. R. . (2017). Hybrid Classifier for Increasing Accuracy of Fitness Data Set. IEEE, 1246–1249.
[36] Nimala, K., & ThamizhArasan, R. (2018). HYBRID DATA MINING APPROACHES FOR ACCURATE PREDICTION OF DIABETES AND HEART DISEASE. International Journal of Pure and Applied Mathematics, 120(6), 2693–2705
[37] Singhal, N., & Ashraf, M. (2015). Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm. International Conference on Computing, Communication & Automation, 138–141. https://doi.org/10.1109/CCAA.2015.7148360
[38] Kaushik, D., & Kaur, K. (2016). Application of Data Mining for High Accuracy Prediction of Breast Tissue Biopsy Results. IEEE Transactions on Knowledge and Data Engineering, 40–45.
[39] Kazienko, P., Lughofer, E., & Trawiński, B. (2011). Hybrid and Ensemble Methods in Machine Learning. New Generation Computing, 29(3), 241–244. https://doi.org/10.1007/s00354-011-0300-3
[40] Castillo, O., Melin, P., & Pedrycz, W. (2007). Hybrid Intelligent Systems: Analysis and Design (Studies in Fuzziness and Soft Computing). Springer, Berlin Heidelberg.
[41] Corchado, E., Abraham, A., & De Carvalho, A. (2010). Hybrid intelligent algorithms and Applications. Information Sciences, 180(14), 2633–2814
[42] Kajdanowicz, T., Kazienko, P., & Kraszewski, J. (2010). Boosting algorithm with sequence-loss cost function for structured prediction. In R. M. Graña, E. Corchado, & S. M. . Garcia (Eds.), Hybrid Artificial Intelligence Systems (pp. 573–580). Berlin, Heidelberg: Springer.
[43] Kuncheva, L. (2004). Combining pattern classifiers: Methods and algorithms. Southern Gate, Chichester, West Sussex, England: John Wiley & Sons.
[44] Sumana, B. V., & Santhanam, T. (2014). An Empirical Comparison of Ensemble and Hybrid Classification. Proc. of Int. Conf. on Recent Trends in Signal Processing, Image Processing and VLSI, 4322–4322. https://doi.org/10.1158/1538-7445.am2015-4322
[45] 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. https://doi.org/10.1109/CIVEMSA.2016.7524248
[46] Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10), 9468–9476. https://doi.org/10.1016/j.eswa.2012.02.112

[47] Jiang, L., & Li, C. (2011). Scaling up the accuracy of decision-tree classifiers: A naive-bayes combination. Journal of Computers, 6(7), 1325–1331. https://doi.org/10.4304/jcp.6.7.1325-1331
[48] WHO. (2019). Malaria. Retrieved November 26, 2019, from WHO website: https://www.who.int/en/news-room/fact-sheets/detail/malaria
[49] World Health Organization[WHO]. (2019). World Malaria Report 2019. Retrieved from https://www.who.int/publications-detail/world-malaria-report-2019
[50] Ogwoka, T. M., Cheruiyot, W., & Okeyo, G. (2015). A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms. International Journal of Computer Applications Technology and Research, 4(9), 693–697. https://doi.org/10.7753/ijcatr0409.1009
[51] Dubey, V. K., & Saxena, A. K. (2016). Hybrid classification model of correlation-based feature selection and support vector machine. 2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2016, 1–6. https://doi.org/10.1109/ICCTAC.2016.7567338
[52] Raghavendra, S., & Indiramma, M. (2016). Hybrid data mining model for the classification and prediction of medical datasets. Int. J. Knowledge Engineering and Soft Data Paradigms, 5(3/4), 262–284.
[53] Ren, Y., Fei, H., Liang, X., Ji, D., & Cheng, M. (2019). A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. BMC Medical Informatics and Decision Making, 19(51). https://doi.org/10.1186/s12911-019-0765-4
[54] Malathi, D., Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., & Sangaiah, K. (2019). Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System. Computers and Electrical Engineering, 73, 114–127. https://doi.org/10.1016/j.compeleceng.2018.11.009
[55] Ju-young, S., Rang, K. K., & Jong-chul, H. (2020). Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management. Agricultural and Forest Meteorology, Vol. 281. https://doi.org/10.1016/j.agrformet.2019.107858
[56] Fan, Junliang, Wu, L., Ma, X., Zhou, H., & Zhang, F. (2020). Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renewable Energy, 145, 2034–2045. https://doi.org/10.1016/j.renene.2019.07.104
[57] Benhar, H., Idri, A., & Fernandez-Aleman, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. In Computer Methods and Programs in Biomedicine. https://doi.org/10.1016/j.cmpb.2020.105635
[58] Aggarwal, C. (2015). Data mining:The Text book. https://doi.org/10.1007/978-3-319-14142-8 14
[59] Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining : An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research, 173, 781–800. https://doi.org/10.1016/j.ejor.2005.07.023
[60] Witten, I., Frank, E., & Hall, M. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.
[61] Maslove, D. M., Podchiyska, T., & Lowe, H. J. (2013). Discretization of continuous features in clinical datasets. 544–553. https://doi.org/10.1136/amiajnl-2012-000929
[62] Li, G., Zhou, X., Liu, J., Chen, Y., Zhang, H., Chen, Y., … Nie, S. (2018). Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Neglected Tropical Diseases, 12(2), 1–19. https://doi.org/10.1371/journal.pntd.0006262
[63] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications, 41, 1937–1946. https://doi.org/10.1016/j.eswa.2013.08.089
[64] Polo, J. L., Berzal, F., & Cubero, J. C. (2007). Weighted Classification Using Decision Trees for Binary Classification Problems. II Congreso Español de Informática, 333–341.
[65] Prasad, N., & Naidu, M. M. (2013). Gain Ratio as Attribute Selection Measure in Elegant Decision Tree to Predict Precipitation. EUROSIM Congress on Modelling and Simulation, 141–150. https://doi.org/10.1109/EUROSIM.2013.35
[66] Hall, M. (2006). A decision tree-based attribute weighting filter for Naive Bayes. In Research and Development in Intelligent Systems XXIII - Proceedings of AI 2006, the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. https://doi.org/10.1007/978-1-84628-663-6-5
[67] Yildirim, P., & Birant, D. (2014). Naive Bayes Classifier for Continuous Variables using Novel Method (NBC4D) and Distributions. IEEE.

[68] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005
[69] Ali, M. F. M., Asklany, S. A., El-wahab, M. A., & Hassan, M. A. (2019). Data Mining Algorithms for Weather Forecast Phenomena : Comparative Study. International Journal of Computer Science and Network Security, 19(9), 76–81.
[70] MAKHTAR, M., NAWANG, H., & SHAMSUDDIN, S. N. W. (2017). Analysis on Students Performance Using Naïve classifier. Journal of Theoretical and Applied Information Technology, 95(16), 3993–4000. Retrieved from www.jatit.org
[71] Zhang, H., Jiang, L., & Yu, L. (2020). Class-specific attribute value weighting for Naive Bayes. Information Sciences, 508, 260–274. https://doi.org/10.1016/j.ins.2019.08.071
[72] Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive bayes text classifiers. Knowledge-Based Systems, 100, 137–144.
[73] Kharya, S., & Soni, S. (2016). Weighted Naive Bayes Classifier : A Predictive Model for Breast Cancer Detection. International Journal of Computer Applications, 133(9), 32–37.
[74] Raschka, S. (2018). Model Evaluation , Model Selection , and Algorithm Selection in Machine Learning. Wisconsin–Madison.
[75] Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Computer Science, 95, 264–271. https://doi.org/10.1016/j.procs.2016.09.332
[76] R Core Team. (2020). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/
[77] Casas, P. (2019). funModeling: Exploratory Data Analysis and Data Preparation Tool-Box. Retrieved from https://cran.r-project.org/package=funModeling
[78] Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. Retrieved from https://cran.r-project.org/package=dplyr
[79] Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data*. R Foundation for Statistical Computing
[80] Kuhn, M. (2020). caret: Classification and Regression Training. Retrieved from https://cran.r-project.org/package=caret
[81] Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/.
[82] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml]
[83] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(16), 1–16. https://doi.org/10.1186/s12911-020-1023-5
[84] El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. International Conference on Communication, Management and Information Technology, 65, 459–468. https://doi.org/10.1016/j.procs.2015.09.132
[85] Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology. https://doi.org/10.1007/s12553-020-00438-1
[86] Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2016). A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. International Journal of Computer Science and Information Security (IJCSIS), 14(12), 868–879. Retrieved from http://arxiv.org/abs/1704.02799
[87] Lee, C. H. (2018). An information-theoretic filter approach for value weighted classification learning in naive bayes. Data & Knowledge Engineering, 113, 116–212.
[88] Yu, L., Jiang, L., Wang, D., & Zhang, L. (2018). Toward naive bayes with attribute value weighting. Neural Computing & Applications.
[89] Zaidi, N. A., Cerquides, J., Carman, M. J., & Webb, G. I. (2013). Alleviating naive bayes attribute independence assumption by attribute weighting. Journal of Machine Learning Research, 14, 1947–1988.
[90] Dinov, I. . (2020). Evaluating Model Performance. Data Science and Predictive Analytics. http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/13_ModelEvaluation.html.
[91] Parikh, R. ., Mathai, A., Parikh, S., Sekhar, C. ., & Thomas, R. . (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Opthamology, 56(1), 45–50

[92] Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. International Conference on Advanced Computing, (6). https://doi.org/10.1109/IACC.2016.25

[93] Wang, Z. (2018). Practical tips for class imbalance in binary classification. https://towardsdatascience.com/practical-tips-for-class-imbalance-in-binary-classification-6ee29bcdb8a7.

 J Health Inform Afr. 2020;7(2):29-46. DOI: 10.12856/JHIA-2020-v7-i2-289