International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

A Hybrid Based Classification and Regression Model for Predicting Diseases Outbreak in Datasets

Leopord Hakizimana[a]*, Prof. Wilson Kipruto Cheruiyot[b], Prof. Stephen Kimani[c], Mlambo Nyararai[d]

[a,b,c,d]*School of Computing and Information Technology (SCIT), Jomo Kenyatta University of Agriculture and Technology*

[d]*University of Rwanda - College of Science and Technology - School of ICT*

[a]*Email: hakizaleo1983@gmail.com*

[b]*Email: wilchery68@gmail.com*

[c]*Email: stephenkimani@googlemail.com*

[d]*Email: nyararaim@yahoo.co.uk*

Abstract

Nowadays, it has been noted that using the application of data mining techniques for predicting the outbreak of the disease has been permitted in the health institutions which have relative opportunities for conducting the treatment of diseases. The main target of this paper is to develop a hybrid based classification and regression model for diseases outbreak prediction in datasets. In this view, the mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression are applied to develop this hybrid model. Accordingly, in hybrid model from this paper there is a core achievements of getting an enhancement as the results described from experiments for combination of more than one Algorithms or methods classifier models discovered that some Algorithms can boost or enhance others through hybrid so that they become more strong significant basing on the accuracy of 100% as output results from hybrid training and with the accuracy of 75% as output results from hybrid evaluation and based on other metrics measurement described on tables 4.1,4.2 and figures 4.1,4.2.

Keywords: Classification; Regression; Hybrid Model; disease outbreak prediction.

----

* Corresponding author.

1. Introduction

Most of human beings seem to want to predict the future. However, it is a natural human desire but "It's hard to make predictions, especially about the future" [10]. Over the past years and today, there has been a rapid technological improvement in Computer Science which has led to the evolution and development of data mining technologies in the health sector for the purposes of hidden pattern discovery such as disease prediction, detection, forecasting which has a paramount in health decision-making. Thus, with the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of health informatics to save time, money and human lives [8].

Normally, early prediction mechanisms are critical in reducing the impact of epidemics and preventing the epidemics from becoming unmanageable by making a rapid response. For example, the cholera epidemic killed over 100,000 people worldwide and sickened 35 million people during the year 2010 (Enserink, 2010). Some researchers estimated that 17 million people die of cardiovascular diseases (CVD) every year [7].

The majority of researchers such as [4] revealed that during 2014 Ebola outbreak in West Africa was the longest, largest and deadliest and the most complex ever witnessed globally. The Ebola virus disease (EVD) epidemic in Guinea, Liberia, and Sierra Leone was the longest, largest, deadliest and the most complex and challenging. Ebola outbreak in history, it was unprecedented in terms of its duration, size of infections and fatality, and geographical spread unlike the past outbreaks which lasted for a very short time, the West African Ebola case has lasted for more than one year and has not yet fully abated. As at 11th February 2015, there were 22,859 EVD cases in total: 3,044 in Guinea, 8,881 in Liberia, and 10,934 in Sierra Leone – with a cumulative death of 9,162 victims.

By explanation, an outbreak or an epidemic is the occurrence of a health-related event (illness, disease complications, and health-related behavior) clearly in excess of the normal expectancy. An epidemic may include any kind of disease, including noninfectious conditions. There is no general rule about the number of cases that must exist for an outbreak to be considered an epidemic. Rather, an epidemic exists when the number of cases exceeds that of what is expected on the basis of past experience for a given population. For example, one case of smallpox would constitute an outbreak. There is no rule on geographic extent. An outbreak could be in only one area or in several countries. When an epidemic spreads in several countries, usually affecting many people, it is called a pandemic. Most flu epidemics that occur during the winter are pandemics. AIDS is considered a pandemic disease. An outbreak may encompass any time period. It may last a few hours (bacterial food poisoning), a few weeks (hepatitis) or several years (acquired immunodeficiency syndrome, or AIDS) [3].

Fortunately, the disease outbreak prediction models are increasingly gaining popularity since these models are developed to predict the disease outbreaks which are becoming common increasingly worldwide. Moreover, medical data mining has a great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis and predictions. Even though the available raw medical data are widely distributed, there is heterogeneous in nature and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system.

Data mining technology provides a user-oriented approach to the novel and discovers hidden patterns in the data [1].

Recently, the huge amount of data being collected and stored in databases or dataset format has recently increased due to the advancements of interest to researchers in data mining, machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for modeling ultimate purposes.

The researcher [2] has discussed in the knowledge discovery and data mining (KDD), he stated that KDD is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. Again the term knowledge discovery in databases or KDD for short refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. Applications of data mining have already been proven to provide benefits to many areas of medicine including diagnosis, prognosis, and treatment.

There are various data mining algorithms for disease predictions such as classification or regression. Despite these models, some are used as a single model. The mixture of classification and regression are possible to establish a hybrid model that can predict the diseases outbreaks on specific datasets reflecting on previously learned information from historical data with accuracy approach.

Therefore, the main purpose of this paper is to develop a hybrid based classification and regression model in the diseases outbreak prediction in datasets where it has been observed that some of the single data mining techniques have the weakness of accuracy. Therefore, this research paper is showing that the hybridization model should overcome the single model's weakness by combining more than one technique where FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression are combined to achieve a Hybrid Based Classification and Regression Model for Predicting Disease Outbreak in Dataset with high accuracy as highlighted and described in the next blocks such as experiment tables 4.1,4.2 and figures 4.1,4.2

2. Dataset

With the respect of developing and testing a hybrid based classification and regression model for diseases outbreak prediction in dataset, for having an effectiveness of the proposed hybrid model, Malaria outbreak Dataset is applied where the dataset is one of the most important factors which can influence the success or failure of outbreak prediction model as cited and used by [15].

The data was collected from different sources like National Vector Borne Disease Control Program, Pune and Meteorological data from Indian Meteorological Department, Pune. Duration of data is from 2011 to 2014. Table 3.1 shows Malaria Outbreak Disease Dataset used for model training and testing and other parameters used such as Average monthly rainfall, Temperature, Humidity, Total number of positive cases, Total number of Plasmodium Falciparum(pF) cases and outbreak occurs in binary values Yes or No

Table 2.1: Malaria Outbreak Disease Dataset

Max.Temperature	**Min.Temperature**	**Avg. Humidity**	**Rainfall**	**Positive**	**pf**	**Outbreak**
29	18	49.74	0	2156	112	No
34	23	83.27	15.22	10717	677	Yes
40	23	50.74	0	1257	127	No
34	24	59.16	9.06	4198	211	No
34	27	73.23	0	11808	712	Yes
31	24	88.77	41.4	10881	648	Yes
33	24	77.94	23.88	8830	459	Yes
31	24	84.57	11.15	9693	482	No
36	24	53.4	2.12	9310	549	No
32	23	57.5	0	13154	838	Yes
34	18	59.4	0	2197	136	No
42	24	49.43	2.19	3362	213	No
45	32	34.74	0.38	416	26	No
43	28	69.07	4.65	7514	410	No
33	23	80.97	6.92	10990	390	Yes
32	24	87.32	11.92	6536	338	No
40	27	63.97	0	11169	776	Yes
39	25	47.52	0	8131	312	No
36	26	72.78	3.54	5138	213	No
31	23	73.35	4.97	10659	612	Yes
30	23	86.81	7.21	9041	418	No
30	22	78.8	3.12	11265	404	Yes
33	22	73.71	1.75	9233	212	No

3. Proposed Model Framework

As the aim of this paper is to develop a hybrid based classification and regression model for diseases outbreak prediction in datasets, the researchers stated that the mining patients' data to predict outbreak diseases or to make a decision by using patients' personal and medical information, it requires steps to be followed [11]. In this research, the steps for proposed model are as shown in figure 3.1. Generally, before building a hybrid model as prediction model and using it for predicting purposes, primarily it is needed to decide which learning algorithm should be used to construct the model. Thus, the predictive performance of the learning scheme should be determined, especially for future data. However, this step is often neglected and the resultant prediction model may not be reliable [6]. The major contribution of this study is to develop an effective hybrid based classification and regression model for predicting outbreak diseases in the dataset. As noted [18] that the literature analysis is a systematic examination of a problem, by means of an analysis of published sources, undertaken with a specific purpose in mind so that in the context of this study, literature analysis was conducted with the two objectives in mind: (i) to develop the hybrid prediction model framework for outbreak diseases, and (ii) to validate the developed model for superiority. The following ways have been done to achieve the above-mentioned objectives (i), literature analysis was conducted in order to identify the existing data mining prediction model framework for outbreak diseases and determine how these models can be collectively used in outbreak diseases prediction. On the other hand, objective (ii) was achieved by performing comprehensive literature analysis for the purpose of establishing a benchmark for evaluating the superiority of the developed framework. Therefore, the framework for this research consists of some phases as detailed in figure 3.1. As the conceptual framework for model proposed.

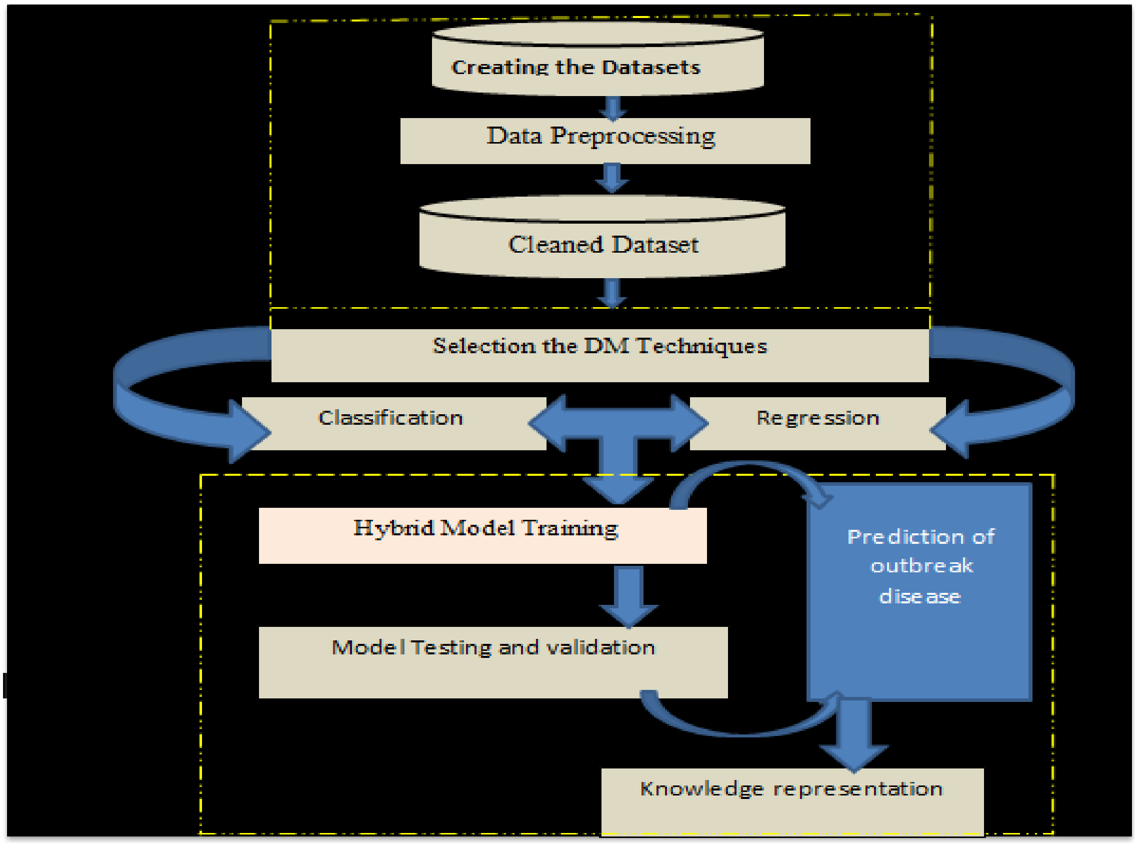


Figure 3.1: Proposed Framework Model Proposed

The steps undertaken to apply data mining algorithms shown in figure 3.1 are presented as follows

i. Creating the Datasets

In the view of the data mining prediction model establishment, the first layer is to obtain real datasets through data collection where the data can be sourced from healthcare organizations such as hospitals, laboratories, and medical centers. For this work, Malaria outbreak Dataset was used and consists of parameters like Average monthly rainfall, Temperature, Humidity, Total number of positive cases, Total number of Plasmodium Falciparum(pF) cases and outbreak occur in binary values Yes or No. This Malaria dataset was cited and used by [15] and from different sources like National Vector Borne Disease Control Program, Pune and Meteorological data from Indian Meteorological Department and Pune with Duration of data from 2011 to 2014.

ii. Data Preprocessing

As noticed by [11] that the second step encloses transforming and preprocessing of data. Data collected from healthcare organizations obtained into one single form understandable by data mining tool. Data comes from different resources each with its own different form. This different form of data needs transformation and preprocessing. This transformation and preprocessing consists of the data cleaning, matching, combining, removing noisy and relevant Data and Standardization. Moreover, in this step involves transforming the data, normalization of data, and removal of noise. In the specified records that should be considered as available records, all tuples were checked to make sure that there are no missing attributes.

iii. Cleaned Dataset

In this layered view, after data preprocessing the data should be kept in special database place for ready to be applied in model training. Also, it is considered as the third step where it consists of data storage process. In data storage, transformed data or cleaned data are stored in a single database with the same format considered as a cleaned dataset. So, these data can be used to apply data mining techniques on.

iv. Selection the DM Techniques for prediction

After data cleaning, the next step is to select the specific algorithms to be applied due to the availability of different techniques such as association, clustering, regression, and classification, etc. But the mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression for regression and classification is considered in this research.

Generally, before elaborating a prediction model and implementing it, there is a need of deciding which learning algorithm should be used to construct the model. Thus, the predictive performance of the learning scheme should be determined, especially for future data usage. However, this step is often neglected and so the resultant prediction model may not be reliable [17]. The algorithms are clarified where the mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression for classification and regression should be considered as the ultimate logarithm in this hybrid model.

v. Hybrid Model Training

In the research[16] author showed that classifiers rely on being trained before they can reliably be used on new data. The more instances the classifier is exposed to the training phase because the more reliable it will be, it will have more experience. However, once trained, we would like to test the classifier too, means that the results accuracy works perfectly.

As clarified in the problem statement, the main objective is to enhance prediction model by applying the amalgamation of more than one algorithm called classification and regression so that there is a mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression.

Furthermore, the hybrid model should be trained using some of classification and regression which is the mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression as known and observed in the other work done. The classification and regression as the mixture of FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression should be chosen as two techniques which should be used to form a real model to predict an outbreak efficiently. Reflecting on most of the successful data mining, machine learning technique is in traditional prediction or forecasting practical application. Also, this layer can consider as Data analysis because, after data storage, next phase is model construction through data analysis. Data analysis is a most important phase in this proposed model. It encloses the following the procedure: First, It includes applying data mining techniques algorithms on patients' data being loaded the dataset into weka data mining tool and computer accuracy and efficiency results of algorithms in terms of producing a stronger algorithm that can be paramount to the doctors, physicians, and health professional etc. To predict diseases outbreak and help them to make decisions for patients' data and to have a spirit of the future trends plans.

vi. Model Testing

This paper[14] presented explained that, in terms of predicting the performance of a classifier on new data, the assessment of its error rate on an independent test set that played no part in the formation of the classifier. The common method of predicting the error rate of a learning technique is to use stratified 10-fold cross-validation. In this view, data should be divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure has executed a total of 10 times on different training sets. Finally, the 10 error estimates are averaged to yield an overall error estimate [14]. For achieving the research objectives, 10-fold cross-validation is paramount, the k-fold cross-validation is used for accuracy, validity, and reliability, where the result of the model is tested using the test dataset and compared based on their accuracy, and the best performing model.

vii. Knowledge representation

After the success of the hybrid model generation named a hybrid of classification and regression, the publication should be done for allowing the user to have insights on it.

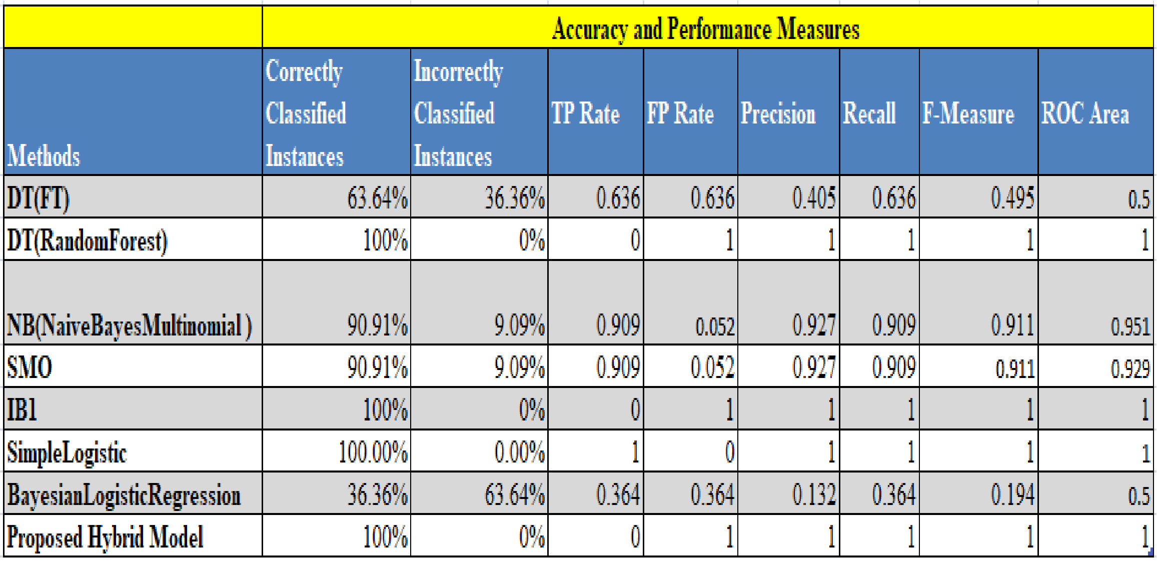
4. Experiments Results

The core objectives of the experiments are to develop and test a hybrid based classification and a regression model for diseases outbreak prediction in datasets. This part section starts by detailing the experiment and results of the hybrid model proposed in this.

4.1 Experiment1: Model Training (Malaria outbreak Dataset)

The experiment of combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression.

Table 4.1: Experiment Performance Evaluation for Malaria outbreak Training Dataset



	Accuracy and Performance Measures							
Methods	Correctly Classified Instances	Incorrectly Classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DT(FT)	63.64%	36.36%	0.636	0.636	0.405	0.636	0.495	0.5
DT(RandomForest)	100%	0%	0	1	1	1	1	1
NB(NaiveBayesMultinomial)	90.91%	9.09%	0.909	0.052	0.927	0.909	0.911	0.951
SMO	90.91%	9.09%	0.909	0.052	0.927	0.909	0.911	0.929
IB1	100%	0%	0	1	1	1	1	1
SimpleLogistic	100.00%	0.00%	1	0	1	1	1	1
BayesianLogisticRegression	36.36%	63.64%	0.364	0.364	0.132	0.364	0.194	0.5
Proposed Hybrid Model	100%	0%	0	1	1	1	1	1

Table 4.1 shows the description of the model's parameter and type of values accuracy and other performance measures results from processed experiment 1 of applying the selected methods and proposed hybrid methods as follows:

FT methods; showing that this classifier performs badly where the classifier accuracy is only 63.64% and also it has TPR 0.636 and FPR 0.636 respectively, precision 0.405, recall 0.636, f-measure0.495 and ROC 0.5 results. This shows the impact of hybridization, the overall results in Table 4.1 indicates that hybrid method has a good and proved obvious improvement in predictive accuracy over some of the single methods.

Random Forest method; showing that it is also good accuracy where it classifies correctly about 100% of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0 and FPR 1 respectively, precision 1, recall 1, f-measure 1 and ROC 1. In Naïve Bayes Multinomial Method(s); showing that it is also good accuracy where it classifies correctly about 90.91 % of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0.909 and FPR 0.052 respectively, precision 0.927, recall 0.909, f-measure 0.911 and ROC 0.951. In SMO classifier; Showing that it is also good accuracy where it classifies correctly about 90.91 % of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0.909 and FPR 0.052 respectively, precision 0.927, recall 0.909, f-measure 0.911 and ROC 0.929. In IB1 Method(s);

Showing that it is also good accuracy where it classifies correctly about 100 % of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0 and FPR 1 respectively, precision 1, recall 1, f-measure 1 and ROC 1. In Simple logistic Method(s); Showing that it is also good accuracy where it classifies correctly about 100 % of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0 and FPR 1 respectively, precision 1, recall 1, f-measure 1 and ROC 1. In Bayesian logistic regression, showing that this classifier performs badly where the classifier accuracy is only 36.36%, so that this shows the impact of hybridization the overall results in Table1 shows that hybrid methods show a good and proved obvious improvement in predictive accuracy over some of the single methods where it has TPR 0.364 and FPR 0.364 respectively, precision 0.132, recall 0.364, f-measure 0.194 and ROC 0.5. In the proposed model named proposed Hybrid which combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression, Showing that it is also good accuracy where it classifies correctly about 100 % of the data; its precision, recall, and f-measure are high as well. And also it has TPR 0 and FPR 1 respectively, precision 1, recall 1, f-measure 1, ROC 1. After receiving the results from training model, it is noticed that the majority of the methods are not good when compared to the hybrid method through voting even if some techniques are good But the weakness of some of them are still an issue that should be boosted or enhanced after being merged with others as indicated on Figure 4.1 describing the Accuracy and Performance Measures of the hybrid proposed model on Experiment I for Training Option using Malaria Dataset.

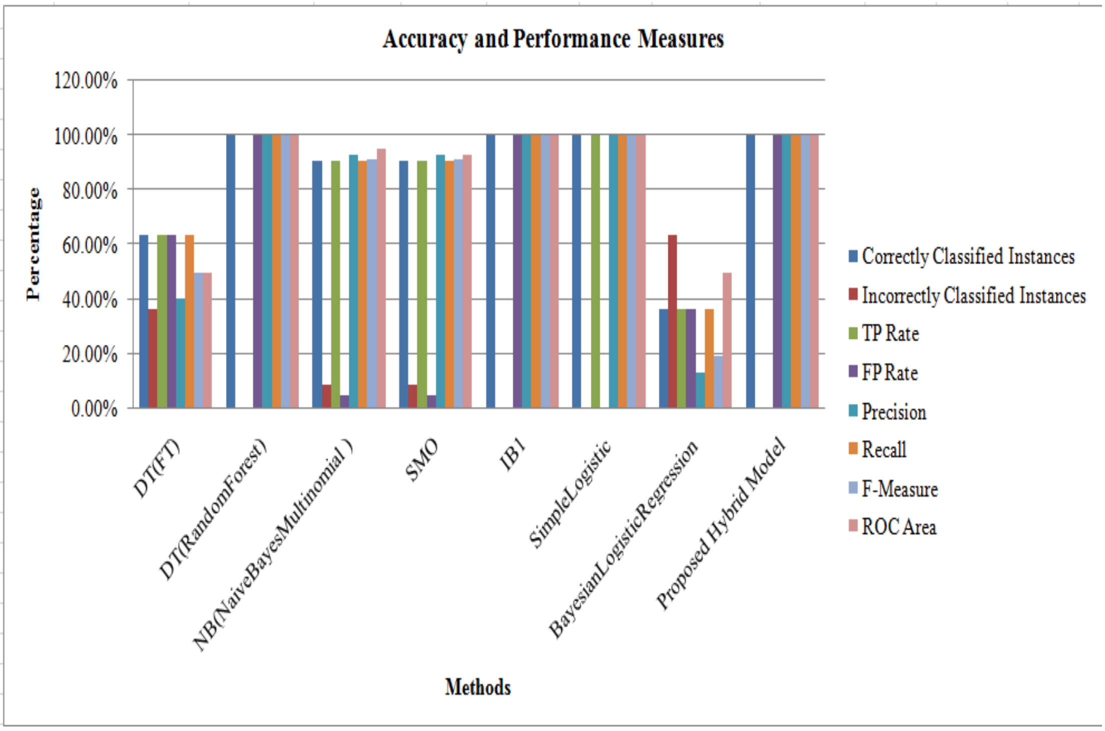


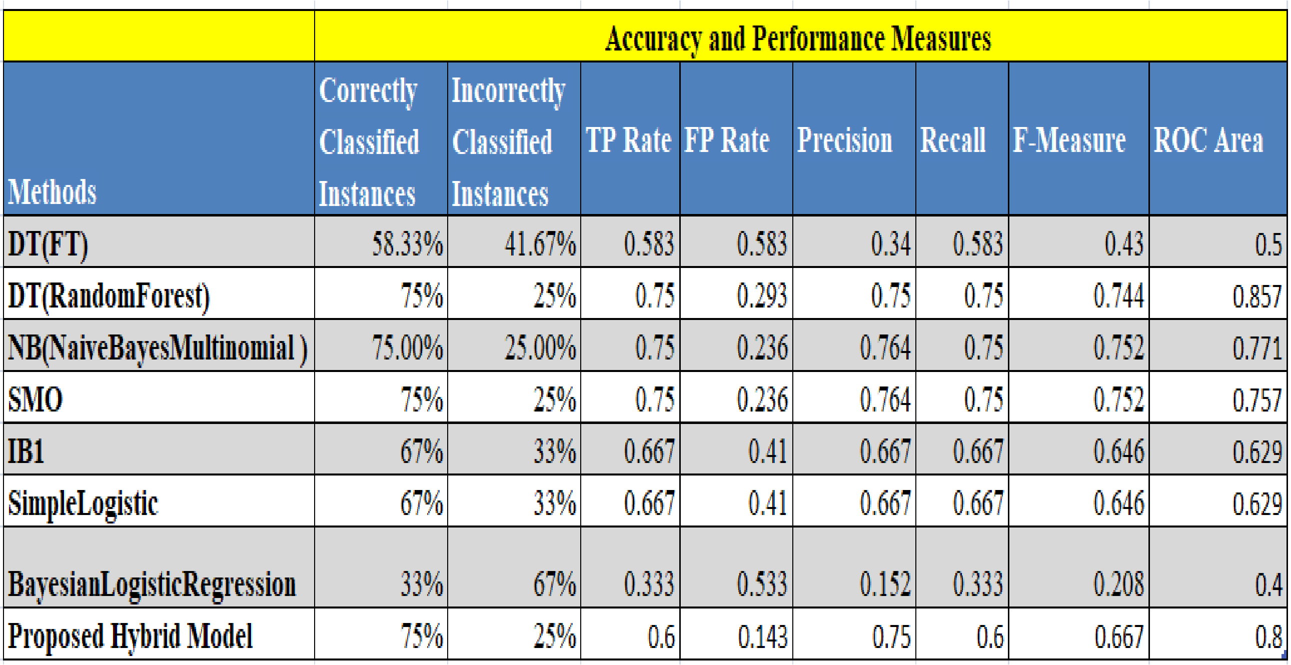
Figure 1.1: Accuracy and Performance Measures on Training Dataset

4.2. Experiment2: Model Evaluation (Malaria outbreak Dataset)

The experiment two carried out on proposed Hybrid which combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic, and Bayesian Logistic Regression algorithms to build a predictive model that named a hybrid-based classification and regression model for predicting diseases outbreak in the dataset, From the Experiment of testing option using Malaria Dataset all attributes were identified to make sound rule and better accuracy due to the WEKA machine learning environment.

This prediction evaluation is done through 10 Fold cross-validation as a testing option using Malaria outbreak Dataset Test, results of the prediction are presented in Table 4.2. The following parameters are selected to describe prediction model classifiers done; the accuracy of the model, sensitivity (TPR), False Positive Rate, F-measure and area under the ROC curve.

Table 4.2: Experiment Performance Evaluation for Malaria outbreak Dataset



	Accuracy and Performance Measures							
Methods	Correctly Classified Instances	Incorrectly Classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DT(FT)	58.33%	41.67%	0.583	0.583	0.34	0.583	0.43	0.5
DT(RandomForest)	75%	25%	0.75	0.293	0.75	0.75	0.744	0.857
NB(NaiveBayesMultinomial)	75.00%	25.00%	0.75	0.236	0.764	0.75	0.752	0.771
SMO	75%	25%	0.75	0.236	0.764	0.75	0.752	0.757
IB1	67%	33%	0.667	0.41	0.667	0.667	0.646	0.629
SimpleLogistic	67%	33%	0.667	0.41	0.667	0.667	0.646	0.629
BayesianLogisticRegression	33%	67%	0.333	0.533	0.152	0.333	0.208	0.4
Proposed Hybrid Model	75%	25%	0.6	0.143	0.75	0.6	0.667	0.8

Table 4.2 considering the above numerical values labeled under each of the evaluation parameters that indicate the model's accuracy and other performance measures results from a processed experiment I of applying the selected models to gain a hybrid model as follows:

FT methods, showing that this classifier performs badly where the classifier accuracy is only 58.33% and also it has TPR 0.583 and FPR 0.583 respectively, precision 0.34, recall 0.583, f-measure 0.43 and ROC 0.5 results so that this shows the impact of hybridization the overall results in Table 4.2 show that hybrid methods show a good and proved obvious improvement in predictive accuracy over some of the single methods.

Random Forest method; showing that it is also good accuracy where it classifies correctly about 75% of the data; its precision, recall, and f-measure are high as well where it has TPR 0.75 and FPR 0.293 respectively, precision 0.75, recall 0.75, f-measure 0.744 and ROC 0.857

In Naïve Bayes Multinomial Method(s); Showing that it is also good accuracy where it classifies correctly about 75% of the data; TPR 0.75 and FPR 0.236 respectively, precision 0.764, recall 0.75, f-measure 0.752 and ROC 0.771 are high as well. In SMO classifier; showing that it is also good accuracy where it classifies correctly about 75% of the data; TPR 0.75, FPR 0.236 respectively, precision 0.764, recall 0.75, f-measure 0.752 and ROC 0.771 as well. In IB1 Method(s); Showing that it is also good accuracy where it classifies correctly about 67 % of the data; TPR 0.667 and FPR 0.41 respectively, precision 0.667, recall 0.667, f-measure 0.646, ROC 0.629 are high as well. In Simple logistic Method(s); Showing that it is also good accuracy where it classifies correctly about 67 % of the data; TPR 0.667 and FPR 0.41 respectively, precision 0.667, recall 0.667, f-measure 0.646, ROC 0.629 are high as well. In Bayesian logistic regression, showing that this classifier performs badly where the classifier accuracy is only 33%, TPR 0.333 and FPR 0.533 respectively, precision 0.152, recall 0.333, f-measure 0.208, ROC 0.4 so that this shows the impact of hybridization the overall results in Table 4.2 shows that hybrid methods show a good and proved obvious improvement in predictive accuracy over some of the single methods. In Hybrid algorithms; it is obvious that the combination of all methods is very accurate and outperformed some of other classifiers, it is classified correctly about 75%, TPR 0.6, FPR 0.143 respectively, precision 0.75, recall 0.6, f-measure 0.667, ROC 0.8 of the testing data as well as it has the highest performance measures. Based on the above results there is a notice that the majority of methods are not good when compared to the hybrid method, But the weakness of some techniques have removed or enhanced after to be merged with others. The overall results in Table 4.2 show that hybrid method has a good and obvious improvement in predictive accuracy over some single methods.

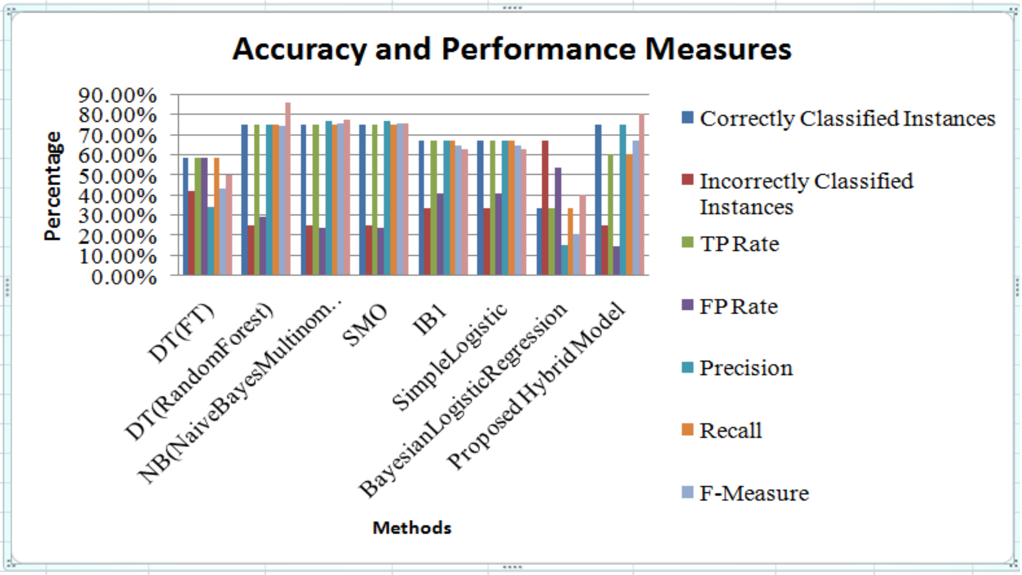


Figure 4.2: Accuracy and Performance Measures on Testing Dataset

Figure 4.2 shows the Accuracy and Performance Measures for proposed Hybrid which combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic, Bayesian Logistic Regression of experiment I through Test Option using Malaria Dataset.

As the results summary shown in figure 4.1 and 4.2 the overall prediction accuracy is improved by a hybrid of classification and regression data mining techniques together as the proposed Hybrid which combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic, Bayesian Logistic Regression. In the view of evaluating the performance of the proposed system, we have tested our approach on different datasets for assuring maxima performance.

In this work Malaria, outbreak Dataset is applied for experimentation. All the experiments are implemented in weka developed in Java language, as Table 4.1, 4.2 shows the results of our proposed hybrid approach and Accuracy and Performance Measures of “Outbreak Disease Prediction hybrid Model” are presented in figure 4.1, 4.2, for two different datasets. From this thesis work, it has been found that our proposed hybrid classifier generates better results in terms of accuracy, sensitivity, and specificity. Therefore, the experimental results describing that our hybrid model is more strong efficient and maxima significance than other techniques of classification and regression of outbreak disease.

According to the related work done, other researchers have been adopted various works to solve the outbreak prediction diseases using single model and same malaria outbreak dataset used by this research, a model was developed with the purpose as an early warning tool to identify potential outbreaks of Malaria where the two popular data mining classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used for Malaria prediction using a large dataset of Maharashtra state. In this regard, the Root Mean Square Error (RMSE) and Receiver Operating Characteristic (ROC) are used to measure the performance of the models. It is observed that performance of the model developed using SVM is more accurate than ANN [15] means that some methods are good than others. Therefore, the hybrid model proposed in this research sounds so good compared to the developed ones, the results are indicating 100% of accuracy on training and 75% of testing.

Several authors including [13] have conducted on Dengue Fever Prediction research where a Data Mining Problem was solved through Naïve Bayesian, REP Tree, Random tree, J48 and SMO techniques, classification techniques to evaluate and compare their performance.

Therefore, each model was reacted as single then the results accuracy performance was compared for looking the good techniques among others. Paper was toward prediction of dengue infection using WEKA Data Mining tool which uses five techniques of classification which are NB, SMO, J48, RT and REP tree, researcher concluded that NB and J48 are the top performance classifier techniques by the way that, they have achieved an accuracy of 92% and 88%, takes few time to run and shows ROC area=0.815, and had smallest error rate.

Due to the numerous existing experiments as the figure 4.5 noted that, Decision tree, Naive Bayes, Neural network and logistic regression were considered as popular data mining algorithms which were used to build the model that predicts whether an individual was being tested for HIV among adults in Ethiopia using EDHS 2011 and so that if you check very well you have found that performance of all of the moles are not similar, where the current research have insight of combination of more that one algorithms for more advanced results [9].

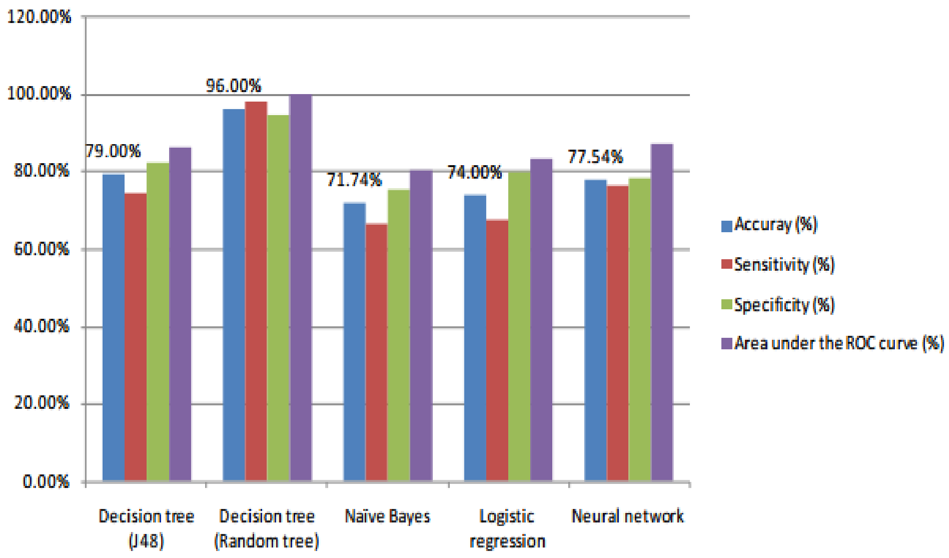


Figure 4.3: Measures of model performance evaluation, Ethiopia, 2013

In this research, with the result noted on Figure 4.3 a shows that single models have different results where the highest is Random tree with 96.00% and lowest is Naïve Bayes with 71.74% . Finally, Based on results tables 4.1, 4.2 and figures 4.1, 4.2 there is a notice that the majority of methods are not good when compared to the hybrid method, but the weakness of some techniques should be removed or enhanced by merging more than one techniques.

5. Conclusions

In this research, a proposed hybrid based classification and regression model for diseases outbreak prediction in datasets were introduced to help a healthcare system firms to deliver good services to its customers by forecasting some outbreak diseases before spreading to more people due to the fact that it is world issue currently. Thus, as computer scientists are now our time to cope with it. Consequently, to achieve the dreams of the study in this research a hybrid model for combining FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression was developed for obtaining a high accuracy of the results of hybrid prediction model on Malaria outbreak disease dataset. Therefore, as reported in the results of the proposed model, it is showing that a hybrid has a good output results than single technique applied, even if some of the single techniques presented a good results but others are weak so that there was a need for combining more than one model, therefore the weakness should be removed or enhanced by amalgamating FT, Random Forest, Naïve Bayes Multinomial, SMO, IB1, Simple Logistic and Bayesian Logistic Regression with the accuracy of 100% as output results from hybrid training and with the accuracy of 75% as output results from hybrid evaluation; hybrid is better based on all metrics measurement described in tables 4.1,4.2 and figures 4.1,4.2.

6. Recommendation for Future Work Directions

This research is not the first or the last one. So, it leaves some research gaps to other researchers for future improvements. Future works may be possible to orient their research or studies on others hybrid classification and regression techniques for the prediction of diseases outbreak due to the availability of classification and regression techniques which not covered in this research.

Better to call upon the next generation researchers to pay attention to the hybrid supervised and unsupervised methods, the next research will be better by performing a deeper analysis into the outbreak diseases. Of course, here a simple dataset was used, so there isn't much space for taking the frequency values into account. But it can be extended to a larger and more comprehensive dataset to analyze the aforementioned values.

Finally, Further studies need to be conducted in this field to gather more information on optimal prediction of outbreak disease which means that it is will be necessary to transform results from Data Mining environment into the business environment and present them in a more comprehensible way such as using programming languages.

Acknowledgement

Special thanks to Prof. Wilson Kipruto Cheruiyot, Prof. Stephen Kimani and my Ph.D. Thesis supervisors who were more than generous with their expertise and precious their time, encouraging, and most of all patients throughout the entire process. I also wish to thank Ferdinand HABIMANA and Noel MANIRAKIZA for their encouragement and guidance.

References

[1] Ujma Ansari, Dipesh Sharma Jyoti Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," International Journal of Computer Applications, pp. 0975 – 8887, 2011.

[2] Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data Richards G, "Data mining for indicators of early," 2001.

[3] Flora Ichiou Huang, "Disease Outbreak Investigation," 2004.

[4] Joséphine Odera, Mr. Benoit Kalasa , Vincent Martin Abdoulaye Mar Dieye, "Socio-Economic Impact of Ebola Virus Disease in West African Countries," Lagos, 2015.

[5] Mirsaeid Hosseini Shirvani Boshra Bahrami, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), vol. 2, no. 2, 2015.

[6] Dulal Chandra Sahana, Software Defect Prediction-Based Classification Rule Mining. Rourkela: Department of Computer Science and Engineering National Institute of Technology Rourkela, 2013.

[7] J., Mensah Mackay, "Atlas of Heart Disease and," Nonserial Publication, 2004.

[8] Shamsher Bahadur Patel, Ashish Kumar Sen Shukla, "A Literature Review in Health Informatics Using Data Mining Techniques," International Journal of software and hardware research engineering, vol. 2, no. 2, 2014.

[9] Tesfay Gidey Hailu, Comparing Data Mining Techniques in HIV Testing Prediction. Addis Ababa, Ethiopia: Intelligent Information Management, 2015.

[10] Hal Kalechofsky, "A Simple Framework for Building Predictive Models," 2016.

[11] Prof. Shailendra Mishra Isha Vashi, "A Comparative Study of Classification Algorithms for Disease Prediction in Health Care," International Journal of Innovative Research in Computer and Communication Engineering, 2016.

[12] Manuel Bayona Flora Ichiou Huang, "Disease Outbreak Investigation," 2004.

[13] Nayyer Masood, Sundas Mehreen, Ulya Azmeen Kamran Shaukat, "Dengue Fever Prediction: A Data Mining Problem," Data Mining Genomics Proteomics, 2015.

[14] Frank Witten, Data mining: practical machine learning tools and techniques., (2nd ed., Ed. San Francisco: Morgan Kaufmann Publishers., 2005.

[15] Ajai Kumar, Lakshmi Panat, Dr. Ganesh Karajkhede, Anuradha Lele Vijeta Sharma, "Malaria Outbreak Prediction Model Using Machine Learning.," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2015.

[16] Roberts. (2005). guide to Weka. http://www.comp.leeds.ac.uk/andyr.

[17] Chandra, S. D. (2013). Software Defect Prediction Based on Classification Rule Mining. Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela { 769 008, India.

[18] Berndtsson, M. &. (2008). Thesis Projects: A Guide for Students in Computer Science and Information Systems. London: Springer-Verlag London.