International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 6, No.2 (Mar-2017)
http://dx.doi.org/10.12785/ijcds/060202

A Novel Holistic Disease Prediction Tool Using Best Fit Data Mining Techniques

Salim A. Diwani[1] and Zaipuna O. Yonah[2]

[1, 2]*Department of Information Technology, Systems Development and Management, Nelson Mandela African Institution of Science and Technology , Arusha. Tanzania.*

Received 7 Nov.2016, Revised 7 Dec. 2016, Accepted 23 Jan. 2017, Published 1 Mar. 2017

Abstract: Given that, today, the healthcare ecosystem is an information rich industry, there is an increasing demand for data mining (DM) tools to improve the quantity and quality of delivered healthcare; especially in handling patients suffering from deadly diseases such as HIV, Breast Cancer, Diabetes, Tuberculosis (TB), Heart diseases and Liver disorder. Given the fatality nature of these diseases when they remain undetected until at advanced stages, there remains a demand for best classifier tools to assist in diagnosing, detecting and treatment of these life-threatening diseases at their early stages. Complementary to this demand is the fact that the healthcare industry today generates large amounts of complex data about patients, hospital resources and disease diagnosis. Consequently, the healthcare ecosystem is warehousing large amount of medical data, which is an asset for healthcare organizations if properly utilized. The large amount of patient and disease related data could be processed and analyzed for knowledge extraction that enables support for cost savings and decision making towards delivery of timely and quality healthcare.
In this paper, we report on an ongoing research work to develop and test a holistic DM disease prediction (Diagnosis and prognosis) tool, equipped with processes for preprocessing patients' data and a learning procedure for selecting a disease-specific best classifier, for disease prediction and delivery of speedy and cost effective diagnostic interventions and patient follow up in a hospital environment. As diseases are diagnosed, the predictive tool helps medical doctors in decision-making about what disease case it is and suggests possible treatment strategies within a much-reduced time. Test results for breast cancer and HIV data sets are reported. Achieved from the reported work are classification accuracies of 97.0752% (Classifier acting singly); 97.6323% (fusion of three classifiers). These results are better than those reported in the literature. The results show that the proposed DM disease prediction tool has potential to greatly impact on current patient management, care and future interventions against deadly diseases.

Keywords*:* Healthcare delivery, Data mining, Electronic medical records, HIV, Breast cancer, Diabetes, Tuberculosis, Heart diseases, and Liver disorder**.**

1. INTRODUCTION

Fundamental to healthcare industry is that health is life, and without good health there is no life. Health is now conceptualized to be beyond the absence of disease to include good diet, sleeping well, physical wellbeing, positive attitudes and social justice. In the context of this paper, healthcare is seen as the attention to the physical health of a human being. Also, healthcare is seen as the treatment, management and prevention of disease and preservation of the physical and mental wellbeing of a person with the help of medical and allied health professionals.

Comparatively, many developing countries like Tanzania, have long suffered from inadequate and poor quality health services largely from lack of skilled health care workers, capable of performing at specialist level in medical disease diagnosis and prognosis. Overall, health sectors in developing countries, also suffer from long waiting times at health facilities for quality health care. For example, many patients spend a lot of time moving from one facility to another seeking diagnostic services. As this happens, patients are usually exposed to too much unnecessary care, lack of transparency, and delays during registration and payment for the healthcare, and avoidable harm to patients. Such waiting times and inconveniences are now known to cause huge economic losses: in human terms and waste in billions of dollars [1], [2]. Essentially, it is a human resource crisis tied to a laboring referral system that is mainly affecting the quality of healthcare delivery. In order to mitigate this challenge there is a need for intervention tools that can assist the healthcare workers in proper disease diagnosis and prognosis towards effective utilization of available human resources. Of interest in this paper, is the use of best fit data mining (DM) techniques to develop and

E-mail: diwania@nm-aist.ac.tz, zaipuna.yonah@nm-aist.ac.tz

implement a disease diagnosis and prediction tool with the goal of improving the quantity and quality of delivered healthcare; especially in handling patients suffering from deadly diseases such as HIV, Breast Cancer, Diabetes, Tuberculosis (TB), Heart diseases and Liver disorder. Given the fatality nature of these diseases when they remain undetected until at advanced stages, there remains a demand for best classifier tools to assist in detecting these life-threatening diseases at their early stages. Complementary to this demand is the fact that the healthcare industry today generates large amounts of complex data about patients, hospital resources and disease diagnosis. Consequently, the healthcare ecosystem is an information rich industry, warehousing large amount of medical data – a kind of "Big Data" - which is an asset for healthcare organizations if properly utilized. The large amount of patient- and disease- related data could be processed and analyzed for knowledge extraction that enables support for cost savings and decision-making towards delivery of timely and quality health care.

In this paper, we report on an ongoing research work to develop and test a holistic data mining (DM) disease prediction tool, equipped with best classifier, for use to deliver speedy intervention and patient follow up in a hospital environment. Thus, better addressing patient's needs, with the potential to improve care quality and to reduce care costs. Disease cases considered include HIV, Breast Cancer, Diabetes, Tuberculosis (TB), Heart diseases and Liver disorder, all in their early stages of development. The paper is organized into six sections. Section II covers a literature review related to the reported work. Section III describes the Methodology used in the prediction design and testing of the prediction tool; and reports on the data collection component and cleaning algorithms of the prediction tool. Section IV reports about disease learning, classification and prediction components of the diagnosis tool and classifier performance criteria. Section V reports on test results from the disease prediction tool for breast cancer and HIV diseases, and Section VI carries the conclusions.

2. LITERATURE REVIEW

Several pioneers in the healthcare industry have attempted to propose interventions for enhancing the delivery of healthcare [3-8]. Kuttikrishnan, *et al*. [3] propose a healthcare system to assist clinicians at the point of care by enabling the clinician to interact with the system to help determine diagnostic and prognosis of patient's data. The system consists of three parts: the knowledge base, inference engine and mechanism to communicate. In [4], Zhou, *et al*. propose a traditional Chinese medicine (TCM) clinical data warehouse (CDW) for medical knowledge discovery and decision support. The TCM is a clinical data warehouse system that incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and TCM clinical decision support system. Karya, [5], proposes data mining techniques for diagnosis and prognosis of breast cancer from Wisconsin database. In the experiment, different classifiers were used and their performance compared; and decision tree classifier was found to be superior with a classification accuracy of 93.62%, followed by Naïve Bayes with a classification accuracy of 84.5%. Sudhir, *et al*. [6], propose a Neural Network Aided Breast Cancer Detection and Diagnosis using Support Vector Machine (SVM). Breast Cancer data set from Wisconsin database was used in the experiment. The highest accuracy achieved by SVM is 96.43%, which can be utilized to support doctors' decision to avoid biopsy. In [7], Palaniapan, *et al*. develop a web based user friendly, scalable, reliable and expandable intelligent heart disease prediction system using data mining techniques namely: decision trees, Naïve Bayes and neural networks. The tool can extract hidden patterns associated with heart disease from a database. The CRISP DM is used to build the disease prediction system. In their experimental results, Naïve Bayes performs better with classification accuracy of 95%, followed by decision tree with classification accuracy of 94.93%, followed by a neural network with classification accuracy of 93.54%.Chaurisa, *et al*. [8], present a performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer diseases. They took advantage of technological advancements to develop prediction models for patients with heart diseases and breast cancer survivability. In the experiment, the breast cancer datasets from Wisconsin were used. The results in this case indicated that the Naïve Bayes performs better than other classifiers with 87.1% accuracy.

What can be said about existing Literature on previous research outcomes is that there is no established standard that guides the data mining efforts of finding classifiers with acceptable high classification accuracies. It is concluded that in order to get better classification accuracy it depends on how one conditions the available data sets before prediction, i.e. preprocessing the data sets before the prediction exercise is done; estimation of baseline performance for the data sets and how the parameters of selected learning algorithm(s) are tuned in order to get best performance.

3. METHODOLOGY AND DATA COLLECTION

In this section we present the methodology used in the reported work and the data collection process for the purpose of obtaining data for training and testing the prediction tool.

A. *Data Processing Stages of the Prediction Tool*

The tool predicts a disease the patient may be having based on symptoms. Figure 1 (stages A-N) summarizes the data processing stages of the prediction tool. It is similar to an expert doctor who asks questions to identify symptoms on the patient. Then, the tool classifies and analyzes the symptoms and other feedback data collected from the patient. The tool may ask a patient to take additional or necessary tests based on adequacy of the patient's symptoms. Then the tool selects the patient's data for disease diagnosis.

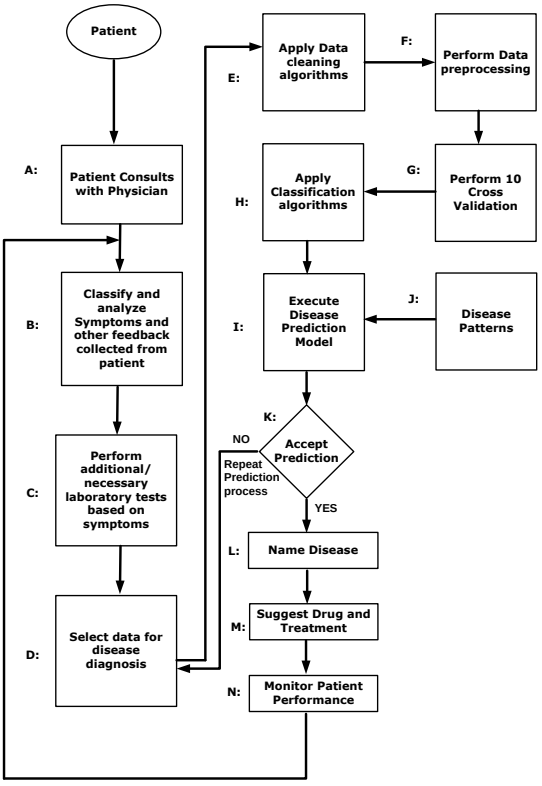

Figure 1. Flow chart of the data mining disease prediction tool.

As illustrated in Fig. 2, data cleaning algorithms are used to clean the selected data. During the data cleaning process, unacceptable values, namely: outliers and extreme values, missing values and noisy data are removed. They are first identified, marked and subsequently handled usually by removing them from the test data sets. These unacceptable values are identified and marked by using appropriate unsupervised attributes depending on the DM platform being used. Equally, the unacceptable instances or samples are removed by using appropriate unsupervised instance filters.

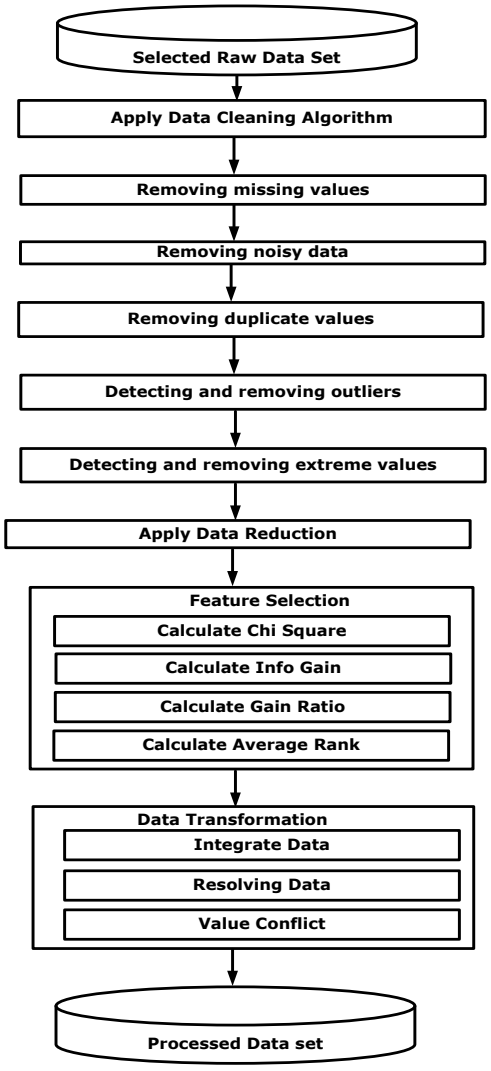

Figure 2. Flow chart of data cleaning algorithm (Process E in Fig.1).

This is done for the purpose of having error free data sets that lead to better classification accuracy from selected classification algorithms during disease prediction. During data pre-processing, some normalization, discretization, data summarization and data reduction techniques are applied. All these procedures are done to make the data sets more appropriate or efficient for the classification task. Lastly, since real data sets are used, often coming from different sources and thus may have different formats, data transformation is done to make sure all have same format.

Then a 10-fold cross validation of the datasets is performed in that learning algorithms are evaluated and compared by dividing data sets into two segments: one segment is used to train the algorithms and the other is

used to validate the algorithms. The training and validation sets must cross over in successive rounds such that each data point has a chance of being validated 10 times against each other and the average is computed. Therefore, the 10 cross validation helps in algorithm selection based on error rate. In the next stage, classification algorithms are applied on the test data sets followed by disease prediction based on the data sets used. Once the prediction is accepted, the disease can be identified. Lastly, the tool will suggest a treatmet strategy and drugs for the patient and thereafter it can be used to monitor the patient's performance.

B. Data Collection (Processes A-D in Figure 1)

The data collection aspect of the prediction tool refers to activities or processes done in stages A-D in Fig. 1. The first source of data contains data sets for breast cancer, diabetes, heart disease and liver disorder available from University of California Irvine (UCI) machine learning data repository. The second source of data contains HIV data sets from Amana hospital in Dar-es-salaam, Tanzania. The data sets are summarized in Table 1. Each data set consists of a number of samples and disease attributes.

TABLE 1. Summary of attributes of data sets of various diseases.

Data Sets	Number of Attributes	Number of Instances	Number of Classes	Attributes
Wisconsin Breast Cancer (UCI Machine Learning Repository)	11	699	2	sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, single epithelial cell size, mitoses, marginal adhesion, bare nuclei, bland chromatin, normal nucleoli, class
CTC Amana Hospital HIV	9	3528	2	weight, nowpregnant, TBScreeningID, ARVStatusCode, ARVCode, CD4, WHOStage, FunctionalStatusCode, class
Pima Indian Diabetes (UCI Machine Learning Repository)	9	768	2	preg, plas, press, skin, insu, mass, pedi, age, class
BUPA Medical Research Liver Disorder (UCI Machine Learning Repository)	6	345	2	age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, class
Cleveland Heart Disease (UCI Machine Learning Repository)	14	303	2	mcv, alkphos, sgpt, sgot, gammagt, class

4. CLASSIFIER TRAINING AND PERFORMANCE EVALUATION

In this study, classifier algorithms to be used in the disease prediction tool were selected due to their popularity and that they are used by many researchers for similar disease classification [9]. Essentially, algorithm selection is a very time consuming task that involves experimentation with different classifiers and analyzing the performance of these classifiers [10]. According to "*No Free Lunch* (NFL) *Theorem*" Duda *et al*. [11], no any single classifier has the best performance in all the data sets. Each data set must be tested using different algorithms or groups of algorithms selected for prediction purposes. In order to get good performance ofthe selected classification algorithms it is crucial to clean the input data. The way data is cleaned has an impact on classifier performance; which depends on how much data can be changed and rearranged. Also, type of variables in the selected data sets are taken into consideration since some algorithms only accept numeric data, some accept only nominal and some algorithms accept both types. Trial and error and Meta learning approaches were used to select the best classification algorithms. In the trial and error approach, available classifiers are applied to the data sets and the best performing ones are selected based on classification accuracy, which ranges from 0 – 100%. Close to 100% an accuracy, the better.

A. Classifier Training

A Meta learning approach aims at automatic discovery of useful algorithms or system. Such a Meta learning process is illustrated in Fig. 3. A database is created with Meta data descriptions of datasets. These meta-data contain estimates of the performance of a set of candidate algorithms on those datasets as well as some Meta features describing their characteristics. A machine learning algorithm is applied to this database to induce the model that relates the value of the Meta features to the performance of the candidate algorithms [12]. The need for Meta learning became necessary given that WEKA data mining tool [13] is being used for algorithms selection. The WEKA platform, as are many DM tools, has many algorithms to select from. This makes the job of selecting algorithms an extremely difficult task.

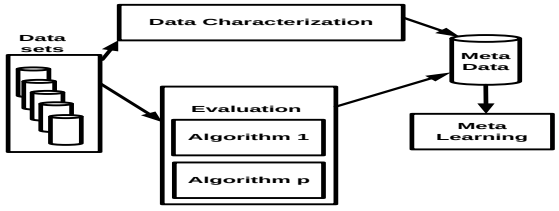

Figure 3. Process of Meta learning [15].

Following the Meta learning exercise, eight different predictive classifiers were selected. These classifiers are: Naïve Bayes (NB), Decision Tree J48 (J48), Instance Based Learning (IBK), Sequential Minimum Optimization (SMO), Multilayer Perceptron (MLP), Decision Tree RepTree, Projective Adaptive Reasonance Theory (PART) and Random Forest (RF) (*individual classifiers*). Given that fusion of such learning algorithms (i.e *combining the algorithms*) can increase the classification accuracy of the resulting fused hybrid classifiers [14], the performance of the eight selected classifiers was investigated by training them to perform classification and prediction while acting singly, in pairs or in groups of three. These classifiers are briefly described in the following paragraphs.

a) Naïve Bayes (NB) - Naives Bayes is a supervised learning algorithms based on Bayes theorem with independent assumption between prediction. Naïve Bayes uses frequency tables built from the data sets used. The Naïve Bayes classfier prediction power is based on the classifier accuracy derived from the concept of probability [16].

b) Decision Tree J48 (J48) - Decision tree is a divide and conquer algorithm, which is a top down approach. The top down approach works by recursively breaking down the complex problem into sub problems and then finding the solutions of sub problems by combining those solutions to form a complex solution. Decision tree uses decision tables built from the data sets used. The core algorithm of decision tree is called ID3 by J.R. Quinlan [17], which uses entropy and information gain to construct a tree. For further details see [18].

c) Instance Based Learning (IBK) - IBK is the K nearest neighbour classification algorithm [19] that is based on similarity functions. K^{th} nearest neighbour is a simple algorithm that predicts new cases of the stored data based on a similarity measure. The prediction is done by selecting the nearest neighbour and calculating which ones are the nearest. The nearest neighbour can be calculated using the Euclidean distance method [20].

d) Sequential Minimum Optimization (SMO) - SMO is a supervised learning algorithm that works the same way as support vector machine (SMV) algorithms that was introduced by Vapnik, *et al*. [21]. SMO is used to classify two different classes and therefore the goal is to design a hyperplane that classifies all training vectors into two classes. The algorithm will select the hyperplane that leaves the maximum margin from both classes and the closest element from those hyperplane. Therefore, the algorithm draws the widest channel or street between two classes.

e) Multilayer Perceptron (MLP) - Multilayer Perceptron (MLP) is a feed forward back propagation network, a very powerful and complicated data mining algorithm based on neurons. MLP is a highly parallel algorithm that processes information much more like a brain rather than a serial computer. MLP is a supervised learning algorithm based on artificial neural network which consists of input layer, output layer and hidden layer. As illustrated in Fig. 4, each layer is made of interconnected nodes; for instance, input layers are interconnected with hidden layers and hidden layers are interconnected with output layers, where the actual processing is done using a system of weighted connections [22], [23].

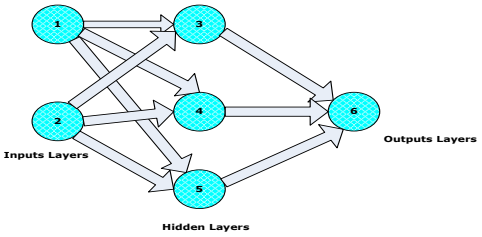

Figure 4. General architecture of MLP.

f) *Random Forest* (RF) - Random forest is a supervised machine-learning algorithm based on decision trees. Random Forest classifier also uses frequency tables processed from data sets. It works the same way as decision tree but this one is more powerful and uses the technique of ensemble learning algorithms by combining weak classifiers to get more powerful classifier. Therefore, random forest works as a large collection of different decision trees algorithms or forest as the name implies all used to make classification. That's the reason random forest uses the bagging technique [24], [25].

g) PART - PART stands for Projective Adaptive Resonance Theory. The inputs to the PART algorithm are the vigilance and distance parameters [26]. PART is a separate-and-conquer rule learner proposed by Eibe and Witten [27]. The algorithm produces sets of rules called decision list, which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART is a partial decision tree algorithms derived from C4.5 decision tree in each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning [28].

h) Decision Tree Rep Tree - Reduced Error Pruning (REP) Tree Classifier is a fast decision tree learning algorithm and builds the tree based on information gain with entropy and minimizes the error arising from variance [29]. This algorithm was first recommended in [30]. REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterwards, it picks the best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and

information gain. Also, this algorithm prunes the tree using reduced-error pruning with back fitting method. At the beginning of the model preparation, it sorts the values of numeric attributes once. As in C4.5 algorithm, this algorithm also deals with the missing values by splitting the corresponding instances into pieces. [31].

B. *Performance Evaluation Criteria*

The selected classifiers were evaluated based on a confusion matrix, which is a visualization tool commonly used to present the accuracy of classifiers [16]. Table 2 sumarizes the entries of the confusion matrix used. The accuracy is a measure of closeness between the target data and the predicted data.

TABLE 2. Confusion matrix for predictive modeling

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	a/(a+b)
	Negative	c	d	Negative Predictive Value	d/(c+d)
	Sensitivity	Specificity		Accuracy = (a+d)/(a+b+c+d)	
	a/(a+c)	d/(b+d)			

The level of performance in the confusion matrix is calculated by identifying the classifier accuracy of correctly classified instances and incorrectly classified instances in terms of percentage of samples tested. The entries in the confussion matrix have the following meaning:

- "*a*"- is *True Positive* (TP), which is the number of positive samples correctly predicted;
- "*b*"- is *False Negative* (FN), which is the number of positive samples wrongly predicted;
- "*c*" - is *False Positive* (FP), which is the number of negative samples wrongly predicted as positive; and
- "*d*" - is *True Negative* (TN), which is number of negative samples correctly predicted.

5. EXPRIMENTAL RESULTS

Two experiments were conducted to evaluate the selected performance of the six learning algorithms (see Section 4).

A. *Learning Experiment Using Wisconsin Breast Cancer Dataset:*

Figure 5 shows performance of the selected classifiers before and after preprocessing input datasets based on a 10-fold cross validation as a test option. Observed is that the accuracy of classifiers is better when preprocessed data are used compared to their performance on unprocessed data; data cleaning makes all the difference. Random Forest performs better with a classification accuracy of 97.0752% compared with other classifiers, followed by SMO and MLP classifiers with accuracies of 96.5181% and 96.3788%, respectively.

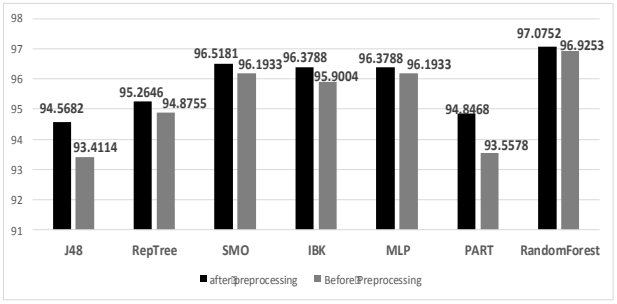

Figure 5. Classification accuracy on breast cancer data set before and after data preprocessing

Figure 6 shows the performance of the classifiers acting in pairs. The RF + SMO combination out performs the other pairs with a classification accuracy of 96.1933%. As a general observation, it is clear that pairing the classifiers does not produce a higher classification accuracy. For example: the RF + RepTree combination performs with accuracy of 95.754% relatively poorer than its individual classifiers, e.g. RF alone has an accuracy of 97.0752%. These results justified the need to consider performance of the classifiers in groups of three. Figure 7 shows performance in terms of classification accuracy of a combination of the SMO + RF with one other classifier. This combination was selected as a primary pair because the two classifiers perform better as a pair (see Fig. 6). Therefore, the classifiers with two other classifiers are SMO+RF+IBK, SMO+RF+J48 and SMO+RF+MLP. It can be observed that the combinations of SMO+RF+IBK and SMO+RF+MLP achieve better performance with classification accuracy of 97.6323%; followed by SMO+RF+J48 with 97.493% accuracy.

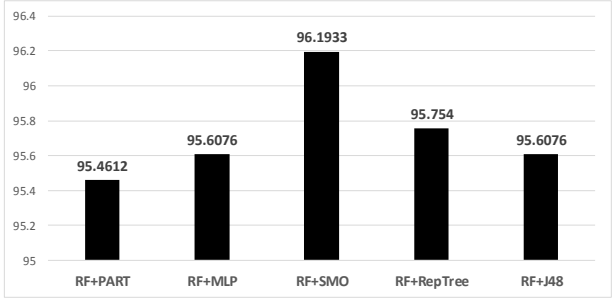

Figure 6. Classification accuracy of fusion of classifiers in pairs after data preprocessing

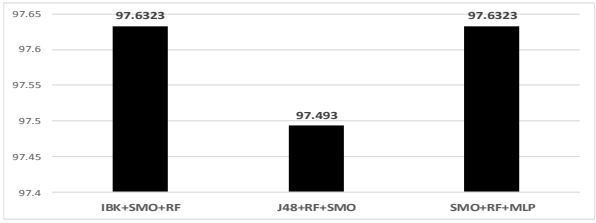

Figure 7. Classification accuracy of fusion of three classifiers after data preprocessing.

After investigating the performance of the selected classifiers acting singly, in pairs and in groups of three, it became necessary to compare their performance in terms of classification accuracies to results reported in the literature. Table 3 summarizes the comparison of accuracies when the classifiers are applied to the same datasets used by other researchers. It can be noted that the new approach produces classification accuracies of 97.0752% (Classifier acting singly); 97.6323% (*fusion of three classifiers*), which are much better than those reported in the literature.

Table 4 shows the prediction results. Out of the available breast cancer data set: 477 samples (known) were used for training, and 206 (unknown) samples were used for testing.

TABLE 3. Comparison of achieved classification accuracies to those reported by other works when using same breast cancer data set.

Method (Reference)	Classifier (s)	Classification Accuracy (%)
[3]	Decision Tree and Naive Bayes	93.62%, 84.5%
[4]	SVM	96.43%
[5]	Naive Bayes and Decision Tree	95%, 94.93%
[6]	Naive Bayes	87.1%
New tool	IBK+SMO+RF, J48+RF+SMO and SMO+RF+MLP	97.6323%, 97.493% and 97.6323%

TABLE 4. Confusion matrix of training and testing data for best classifiers.

Algorithm	Training (477)			Testing (206)		
	Desired Result	Output Result		Desired Result	Output Result	
		Benign	Malignant		Benign	Malignant
SMO	Benign	302	8	Benign	131	3
	Malignant	8	159	Malignant	4	68
RF	Benign	301	9	Benign	130	4
	Malignant	5	162	Malignant	3	69
IBK+SMO+RF	Benign	302	8	Benign	131	3
	Malignant	4	163	Malignant	5	67
J48+RF+SMO	Benign	302	8	Benign	130	4
	Malignant	5	162	Malignant	3	69
SMO+RF+MLP	Benign	302	8	Benign	131	3
	Malignant	7	160	Malignant	4	68

Table 5 shows the sensitivity, specificity and prediction for training and testing data for testing of Breast Cancer data set. Observable is that the combinations: IBK+SMO+RF, J48+RF+SMO and SMO+RF+MLP perform better with sensitivity (0.9869, 0.9837 and 0.9773) and Specificity (0.9532, 0.9529 and 0.9524).

TABLE 5. Performance of training and testing data, with prediction done based on confusion matrix of each classifier

Algorithm	Training (477)				Testing (266)			
	Sensitivity	Specificity	Positive Prediction	Negative Prediction	Sensitivity	Specificity	Positive Prediction	Negative Prediction
SMO	0.9742	0.9521	0.9742	0.9521	0.9704	0.9577	0.9776	0.9444
RF	0.9837	0.9474	0.9709	0.9701	0.9774	0.9452	0.9701	0.9583
IBK+SMO+RF	0.9869	0.9532	0.9742	0.9760	0.9632	0.9571	0.9776	0.9305
J48+RF+SMO	0.9837	0.9529	0.9742	0.9701	0.9774	0.9452	0.9701	0.9583
SMO+RF+MLP	0.9773	0.9524	0.9742	0.9581	0.9704	0.9577	0.9776	0.9444

Table 6 shows the feature selection results of the Breast Cancer attributes. Feature selection is a process of selecting input variables or attributes and then highlighting the importance of attributes that are selected for use in the disease diagnosis and prediction. The main mission of the feature selection is to improve the performance of a classifier. From this table, it can be seen that the best attribute selected was Bare Nuclei followed by uniformity of cell shape and the least attribute selected is mitoses followed by single epithelial cell size.

TABLE 6. Feature Selection results.

Variable	Chi-Square	Info Gain	Gain Ratio	Average Rank	Importance
Clump Thickness	315.52	0.405	0.282	105.40233	5
Cell Size Uniformity	374.317	0.467	0.436	125.07333	3
Cell Shape Uniformity	375.317	0.467	0.436	125.62933	2
Marginal Adhesion	247.246	0.297	0.295	82.612667	7
Single Epithelial Cell Size	236.134	0.277	0.295	78.902	8
Bare Nuclei	377.087	0.465	0.423	125.99167	1
Bland Chromatin	324.629	0.398	0.381	108.46933	4
Normal Nucleoli	268.28	0.319	0.324	89.641	6
Mitosis	57.204	0.067	0.18	19.15033	9

B. Experiment Using CTC HIV Datasets From Amana Hospital:

Figure 8 shows performance of different classifiers before and after preprocessing based on 10-fold cross validation as a test option. Once again, it is demonstrated that classifiers acting on pre-processed datasets achieve better performance in terms of classification accuracy than before preprocessing due to data cleaning. In this HIV case, the SMO classifier performs better with an accuracy of 90.9014% compared with other classifiers, followed by Naïve Bayes and MLP classifiers with classification accuracies of 90.7029% and 90.4762%, respectively.

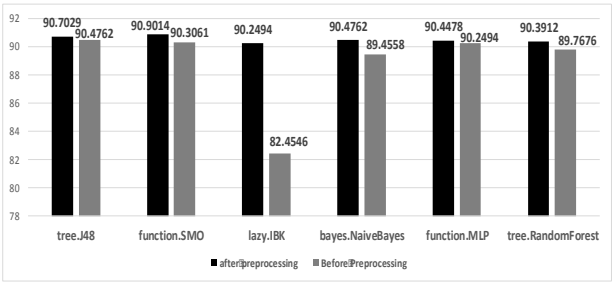

Figure 8. Classification accuracy on HIV data set before and after data preprocessing.

Figure 9 shows the performance of the selected classifiers when acting in pairs. In this case, the SMO + J48 combination out performs the other pairs with a classification accuracy of 90.9297%. Coincidently, the pair slightly performs better than the individual classifiers as in Fig. 8. Again, these results justified the need to consider performance of the classifiers in groups of three.

Figure 10 shows performance in terms of classification accuracy of a combination of the SMO and J48 with one other classifier. This combination is selected as a primary pair because the two classifiers perform better as a pair (see Fig. 9). Therefore, the hybrid classifiers (*classifiers grouped in three*) are SMO+J48+NB and SMO+J48+MLP. It can be observed that the combination of SMO + J48 + MLP achieves better performance with classification accuracy of 91.3265%; followed by SMO+J48 +NB with 91.0998% accuracy. For this reason, the SMO+J48+MLP combination was selected for prediction of HIV disease for the data set from Amana Hospital Dar-es-salaam, in Tanzania.

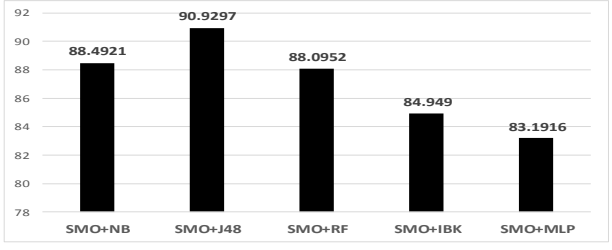

Figure 9. Classification accuracy of classifiers grouped in two after data preprocessing.

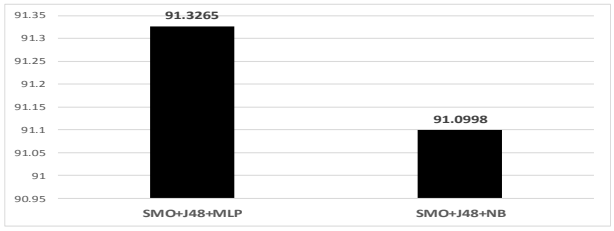

Figure 10. Classification accuracy of classifiers grouped in three after data preprocessing.

Table 7 shows the prediction results. Out of the available HIV data set: 2116 samples (known) were used for training, and 1412 (unknown) samples were used for testing.Table 8 shows the sensitivity, specificity and prediction for training and testing data.

TABLE 7. Confusion matrix of training and testing data.

Algorithm	Training (2116)			Testing (1412)		
	Desired Result	Output Result		Desired Result	Output Result	
		Good	Poor		Good	Poor
SMO	Good	1692	1	Good	1127	3
	Poor	198	225	Poor	119	163
J48	Good	1692	1	Good	1125	5
	Poor	201	222	Poor	121	161
SMO+J48+MLP	Good	1692	1	Good	1125	5
	Poor	200	223	Poor	119	163
SMO+J48+NB	Good	1692	1	Good	1126	4
	Poor	198	225	Poor	119	163

For testing of HIV data set, it was observed that sensitivity for testing dataset performs better than the training dataset. Table 9 shows the feature selection results of the HIV attributes. From the table we can see the best attribute selected is WHOStage followed by CD4 and the least attribute selected is ARVStatusCode followed by NowPregnant.

TABLE 8. Performance of training and testing data; prediction done based on confusion matrix of each classifier.

Algorithm	Training (2116)				Testing (1412)			
	Sensitivity	Specificity	Positive Prediction	Negative Prediction	Sensitivity	Specificity	Positive Prediction	Negative Prediction
SMO	0.8952	0.9956	0.9994	0.5319	0.9045	0.9819	0.9973	0.5780
J48	0.8938	0.9955	0.9994	0.5248	0.9028	0.9699	0.9956	0.5709
SMO+J48+MLP	0.8943	0.9955	0.9994	0.5272	0.9043	0.9702	0.9956	0.5780
SMO+J48+NB	0.8952	0.9956	0.9994	0.5319	0.9044	0.9760	0.9965	0.5780

TABLE 9. Feature selection results.

Variable	Chi-Square	Info Gain	Gain Ratio	Average Rank	Importance
Weight	29.95	0.001717	0.000518	9.98408	6
NowPregnant	7.907	0.000454	0.000487	2.63598	7
TBScreeningID	89.281	0.005225	0.026071	29.7708	5
ARVStatusCode	0	0	0	0	8
ARVCode	1496.627	0.084739	0.040072	498.9173	4
CD4	1574.497	0.088491	0.044713	524.8767	2
WHOStage	1680.81	0.093555	0.063314	560.3223	1
FunctionalStatusCode	1516.893	0.09395	0.200477	505.729	3

6. CONCLUSION

Disease detection and its treatment methods is a major area of concern that needs much attention these days. This paper supports the fact that machine learning can be of big help when it comes to medical diagnosis and prognosis. Presented in this paper is an approach for detection and prediction of two deadly diseases using machine learning techiques. The presented tool can assist physians either new or experienced in medical diagnosis and prognosis at initial stages of the diseases. The main issue here is to save time, reduce healthcare costs, quality healthcare delivery and reduce mortality and morbidity rate, which is very crucial in life threatening diseases.Therefore, the developed tool can help physicians make more accurate diagnosis as well as get answers they often seek from individual patients. As diseases are diagnosed, the predictive tool helps medical doctors in decision-making about what disease case it is and suggests possible treatment strategies within a much-reduced time. Test results for breast cancer and HIV data sets are reported. Achieved in the reported work are classification accuracies of 97.0752%(Classifier acting singly); 97.6323% (fusion of three classifiers). The results show that the fusion of three classifiers is superior to others classifiers [3-6] reported in the literature. Therefore, the results confirm that the proposed DM disease prediction tool has potential to greatly impact on current patient management, care and future interventions against deadly diseases such as HIV, Breast Cancer, Diabetes, Tuberculosis (TB), Heart Diseases and Liver disorder.

REFERENCES

[1] Institute of Medicine, To Err Is Human: Building a Safer Health Care System, Washington, D.C.: National Academies Press, 2000.

[2] B. Jennings, M. A. Baily, M. Bottrell and J. Lynn "Health Care Quality Improvement: Ethical and Regulatory Issues" The Hastings Center Garrison, New York 2007.

[3] M. Kuttikrishnan, I. Jeyaraman and M. Dhanabalachandran., "An Optimised Intellectual Agent Based Secure Decision System for Health Care". International Journal of Engineering Science and Technology, Vol. 2, No. 8, pp. 3662-3675, 2010.

[4] X. Zhou, S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li, Y. Guo, H. Zhang, Z. Gao and X. Yan., "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support". Elsevier Artificial Intelligence in Medicine Vol. 48, pp. 139–152, 2010.

[5] S. Karya., "Using Data Mining Techniques for diagnosis and prognosis of cancer disease". International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[6] S. D. Sawarkar, A. A. Ghatol and A. P. Pande., "Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine". Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, pp. 158-163, 2006.

[7] S. Palaniappan and R. Awang., "Intelligent Heart Disease Prediction System Using Data Mining Techniques". International Journal of Computer Science and Network Security (IJCSNS), Vol.8 No.8, August 2008.

[8] V. Chaurasia and S. Pal., "Performance Analysis of Data Mining algorithms for diagnosis and prediction of Heart and Breast Cancer disease". Review of Research Journal Vol. 3, No. 8, 2014.

[9] J. Jojan and A.Srivihok., "Duo Bundling Algorithms for Data Preprocessing: Case Study of Breast Cancer Data Prediction" Lecture Notes on Software Engineering, Vol. 2, No. 4, November 2014.

[10] S. Cacoveanu, C.Vidrighin and R. Potolea, "Evolution Meta-Learning Framework For automatic Classifier Selection", 2005.

[11] R.O.Duda, P.E. Hart and D. Stork, Patter Classification. J.Wiley & Sons, New York, 2001.

[12] B. C. R. Prudencio and T. B. Ludermin., "Uncertainty Sampling Methods for Selecting Datasets in Active Meta Learning". Proceedings of International joint Conference on Neural Networks, San Jose, California, USA, July 31-August 5, 1082-1089, 2011.

[13] M. Hall, E. Frank, G. Holmes and B. Pfahringer., "The WEKA Data Mining Software: An Update". SIGKDD Explorations, Vol. 11, No. 1, 2009.

[14] X. Ceamanosa, B. Waskeb, J. A. Benediktssonc, J. Chanussotd, M. Fauvele and J. R. Sveinsson., "A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data" International Journal of Image and Data Fusion Vol. 1, No. 4, pp. 293–307, December 2010.

[15] N. Bhatt, A. Thakkar and A. Ganatra., "A Survey & Current Research Challenges in Meta Learning Approaches based on Dataset Characteristics". International Journal of Soft Computing and Engineering (IJSCE)ISSN:, Vol. 2, No. 1, pp. 2231-2307, March 2012.

[16] J. Han and M. Kamber, Data Mining: Concepts and Techniques, CA:Morgan Kaufmann Publishers, 2001.

[17] J.R. Quinlan, Decision trees and multi-valued attributes. In J.E. Hayes & D. Michie (Eds.), Machine intelligence 11. Oxford University Press (in press), 1985.

[18] N. Soonthornphisaj, Artificial Intelligence, Bangkok, Thailand: Kasetsart University, 2009.

[19] A. Christobel and Y, Sivaprakasam.,"An Empirical Comparison of Data Mining Classification Methods". International Journal of Computer Information Systems, Vol. 3, No. 2, 2011.

[20] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal.,"Locating the Nodes: Cooperative Localization in Wireless Sensor Networks". IEEE Signal Process. Mag., Vol. 22, No. 4, pp. 54–69, Jul. 2005.

[21] V.N. Vapnik.,"The Nature of Statistical Learning Theory". 1st ed., Springer-Verlag, New York, 1995

[22] M. C. Popescu, V.E. Balas, L. Perescu-Popescu and N. Mastorakis., "Multilayer Perceptron and Neural Networks". WSEAS Transactions on Circuits and Systems, Vol. 8, No. 7, pp. 579 – 588, Jul. 2009.

[23] E. Mattar, "A Practical Neuro-fuzzy Mapping and Control for a 2 DOF Robotic Arm System**",** *International Journal of Computing and Digital Systems (IJCDS)*, Vol. 2, No. 3, pp. 109-121, 2013.

[24] L. Breiman., "Random Forests". Machine Learning, Kluwer Academic Publishers. Manufactured in The Netherlands, Vol. 45. pp. 5 -32, 2001.

[25] M. Al-Emran, "Hierarchical Reinforcement Learning: A Survey". *International Journal of Computing and Digital Systems (IJCDS)*, Vol. 4, No.2, pp. 137 -143, Apr-2015.

[26] Yongqiang Cao and Jianhong Wu, "Projective ART for clustering data sets in high dimensional spaces", Elsevier Science Ltd, Neural Networks Vol. 15, pp. 105-120, 2002.

[27] Witten IH, and Frank E.. Data mining: practical machine learning tools and techniques – 2nd ed. the United States of America, Morgan Kaufmann series in data management systems., 2005.

[28] Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing 6.2: pp. 119-138, 2006

[29] J. Quinlan, Simplifying decision trees, International Journal of Man Machine Studies, Vol.27, No. 3, pp. 221–234, 1987.

[30] S.K. Jayanthi and S.Sasikala. 2013. REPTree Classifier for indentifying Link Spam in Web Search Engines. IJSC, Volume 3, Issue 2, (Jan 2013), 498 – 505.

[31] J. Kittler, M. Hatef, R. Duin and J. Matas. "On Combining Classifiers". IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 20, NO. 3, March 1998.

Salim Amour Diwani received his BS and M.Sc. degrees in computer science from Jamia Hamdard University, New Delhi, India in 2006 and in 2008, respectively. He is currently a PhD scholar in Information communication Science and Engineering at Nelson Mandela African Institution of Science and Technology in Arusha, Tanzania. His primary research interests are in the areas of Big Data, Data Mining, Machine Learning and Database Management Systems.

Zaipuna O. Yonah -MIET, MIEEE - holds a B.Sc. degree (with Honors - 1985) in Electrical Engineering from University of Dar es Salaam - Tanzania; and M.Sc. (1986) and PhD (1994) Degrees in Computer-based Instrumentation and Control Engineering from the University of Saskatchewan, Saskatoon-Canada. In Tanzania, he is a Registered Consulting Engineer in ICTs.

Yonah has over 31 years of practice. His work spans the academia, industry and policy making fields. He is currently associated with The Nelson Mandela Institute of Science and Technology, Applied Engineering & ByteWorks (T) Ltd and the Institute of Electrical and Electronics (IEEE) Inc. He is one of the mentors and pioneers driving the national broadband agenda in Tanzania and EAC, SADC regions. He believes that ICTs, as tools for development, promise so much: *interactivity*, *permanent availability*, *global reach*, *reduced per unit transaction costs*, *creates increased productivity and value*, *jobs and wealth*, *multiple source of information and knowledge*. Armed with such a belief, his current work aims at creating and delivering value through ICT-enabled services in the shortest times possible. His research interests include: ICT4D, Mobile and Web applications, Big Data, Data Mining, High Performance Computing, high-capacity Broadband networks, Intelligent Instrumentation and Control Engineering and ICT enabled 21st Century Education delivery (ICT4E): *Personalized, Facilitated, and Connected Learning*.