Performance Analysis of Algorithms on Different Types of Health Related Datasets

N N Khanom\textsuperscript{1}, F Nihar\textsuperscript{2}, S S Hassan\textsuperscript{3} and L Islam\textsuperscript{4}

\textsuperscript{1}Department of Computer Science & Engineering, East West University, Bangladesh
\textsuperscript{2}Department of Computer Science & Engineering, East West University, Bangladesh
\textsuperscript{3}Department of Computer Science & Engineering, East West University, Bangladesh
\textsuperscript{4}Department of Computer Science & Engineering, East West University, Bangladesh

Abstract. Healthcare related data are very important for people of the global community to make them aware of their lives and rights. By analyzing performances of different types of health-related data produce efficient valuable rich information; which can be used in further healthcare research. From the previous works we have seen, they have also tried to analyze the performances of algorithms to modify the accuracy but this study is unique as it has used lots of different types of health-related datasets and algorithms together to achieve better accuracy. This study aims to analyze performances of algorithms on different types of healthcare related data to produce effective information in order to assist global community in their daily life. This will help people to lead a healthy and comfortable life and innovate new ideas to further change their lifestyle. The effective information or the standard used here is the accuracy of the applied algorithms.

1. Introduction
Algorithm analysis plays an important role in computer Science field by providing theoretical guidance to solve a particular problem. By analyzing the algorithms, one can determine the time and memory space required for execution. We have analyzed the performance of four different type of algorithms (K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Random Forest) on four different types of datasets. The datasets include Breast Cancer Wisconsin (Diagnostic) [1], Pima Indians Diabetes Database [2], Heart Disease UCI [3], Toddler Autism Spectrum Disorder (ASD) Dataset [4].

2. Related Work
In paper [5], the authors have introduced a paper on “Breast cancer diagnosis via data mining: performance analysis of seven different algorithms” that helps to diagnose breast cancer dataset which was given by the patients. In this paper, they predicted and found accuracy of this dataset by applying seven different algorithms; which are Discriminant Analysis, Naïve Bayes, Artificial Neural Networks (also named as Multi-Layer Perceptron), K-nearest Neighborhood, Support Vector Machine, Logistic Regression, and Decision Trees.

In paper [6], the authors have presented a paper on “Comparative analysis of intelligent hybrid systems for detection of Pima Indian diabetes” that defines various new efficient combined approaches in Biomedicals field. After comparing the various approaches, they also compared their performances. (Evolutionary Artificial Neural Networks, Adaptive Neuro Fuzzy Interface Systems, and Ensembles) and standard back propagation algorithm are three main combined systems that were selected.

In paper [7], the researchers have introduced a paper on “Liver disease prediction using SVM and Naïve Bayes algorithms” that predicts liver diseases using classifier techniques (Support Vector
Based on the performance criterions such as execution time and classification accuracy, these classifier algorithms are compared. According to this paper, SVM is an effective classification technique to predict the liver diseases.

In [8], they have presented a paper on “Heart disease diagnosis using predictive data mining” that depicts and establishes prediction and diagnosis system based on predictive mining for heart diseases. Lots of experiments have been conducted to analyze the performance of different types of techniques including Naïve Bayes and Decision Tree. They have proposed a clinical database of 13-attribute structure. According to this paper, when compared to Decision Tree, Naïve Bayes outperforms.

In paper [9], the author has introduced a paper on “Liver disease prediction using Bayesian classification” that predicts three main liver diseases such as Liver Cancer, Hepatitis, and Cirrhosis with distinct symptoms by applying Bayesian Classification Technique.

In [10], the researchers have introduced a paper on “Performance analysis of various data mining techniques in the prediction of heart disease” that compares accuracy of the classifiers and finds the best classified instances. Based on some factors which are responsible for this disease (family history, obesity, age and others), different classification algorithms are used to anticipate heart disease.

In [11], the authors have proposed a paper on “A critical study of selected classification algorithms for liver disease diagnosis” that classifies the chosen classification techniques for some liver patient datasets. The classifier techniques used here are Back propagation Neural Network, Naïve Bayes classifier and Support Vector Machine. These techniques are analyzed based on precision, accuracy, specificity and sensitivity.

In paper [12], the author has introduced a paper on “Performance analysis of data mining algorithms for breast cancer cell detection” using Decision Tree, Logistic Regression, and Naïve Bayes that detects the smallest subgroup of features that can classify the type of breast cancer (either benign or malignant) accurately. Then a study on different cancer classifier approaches such as Decision Tree, Naïve Bayes, and Logistic Regression classifiers are conducted and here the time complexity of each of the classifier technique is also calculated. Here, Logistic Regression Classifier technique is concluded as the best classification technique with the highest accuracy as compared to the other two classifiers.

In paper [13], the authors have proposed a paper on “Design of a hybrid system for the diabetes and heart disease” that uses crisp and fuzzy values in data and presents a new method. Also, they have proposed a combined neural network that involves fuzzy neural network and artificial neural network. They have used datasets of Pima Indians Diabetes and Cleveland Heart Disease. To analyze the performance, sensitivity, specificity, and accuracy have been used.

In [14], the authors have presented a paper on “Performance analysis of various data mining classification techniques on healthcare data” that aims to analyze the performance of different classification algorithms by using three different machine learning techniques. They have used four different datasets. The standard used here are error rate of every applied classification technique and percentage of accuracy. The experiments are concluded using the ten-fold cross validation approach.

In paper [15], the authors have introduced a paper on “ASD in Infants and Toddlers” that defines symptoms of ASD in early childhood, early diagnosis, and symptoms of Autism Spectrum Disorder in the second and third years of life, prodromal symptoms of Autism Spectrum Disorder in the first year of life including the risks.

In paper [16], the authors have presented a paper on “Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD)” that selects Least Absolute Shrinkage, Chi-square and Selection Operator (LASSO) as feature selection methods in order to choose the most significant three supervised algorithms (K-Nearest Neighbors, Random Forest, and Logistic Regression with K-fold cross validation. Comparing with the others, Logistic Regression gives the best output (97.541%) and this output has been achieved by using 13 selected features on the basis of Chi-square selection method.

3. Methodology

3.1. Decision Tree (DT)

For solving both classification and regression problems, decision tree can be used; where it represents
Tree structure to solve a particular problem. Here, each leaf node represents attributes and a class label; which are represented in the internal structure of the tree. This algorithm has the ability to handle binary and multiclass classification problem.

Algorithm:

Step 1: In this tree there is a split on the training data on the basis of the feature which produces highest information gain. We need two different entropy for calculating. One is for total database that means for all the attribute and the equation for that is, Info(F) = - Σ i log2( ) =1; where pi is the probability that an arbitrary tuple in F belongs to class Yes or No and another is for each attribute and equation for that is, info att(F) = Σ| j|| |× ( ) =1.

Step 2: After that total gain will be calculated. Here, the formula, Gain(attribute) = info(F) – info att(F)

Step 3: Finally, with the gain information, we have the best attribute to choose the root node. After selecting the best attribute with the highest information gain, divide the data set by the branch of this selected attribute and repeat the procedure.

3.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors technique is one of the most effective algorithms to experiment in data mining field, intrusion detection area, and pattern recognition field.

Algorithm:

Here, m is considered as the number of samples in case of training; where, p is an unseen point.

Step 1: The training samples should be stored in an array.

Step 2: for i starting at 0 to m:

   Euclidean distance should be calculated.

Step 3: A set of smallest distances should be built; where each of the distances correspond to an already classified data point.

Step 4: The majority label among the dataset should be returned.

3.3. Support Vector Machine (SVM)

The idea of maximization of the minimum distance from separating hyperplane to the nearest is the main idea of this technique. Basically, for solving the pattern recognition problem, Support Vector Machine is introduced. The basic SVM supports binary classification and extension supports multiclass classification.

Algorithm:

Step 1: In n-dimensional space, each data item is shown using a point; where n is considered as the number of features. Each feature’s value becomes a particular coordinate’s value; which is also included here.

Step 2: Classification techniques should be performed to find out the hyper-plane that differentiates the two classes accurately.

3.4. Random Forest (RF)

Random Forest can perform both classification and regression tasks by using a technique named bagging and multiple decision trees. This technique integrates multiple decision trees to determine the final result; where it does not rely on the decision trees individually. Random forest is responsible to create decision trees based on selecting data samples randomly, it gets accurate prediction from each of the tree and tries to selects the best solution using voting procedure.

Algorithm:

Step 1: At first, pick data points randomly from the training set.

Step 2: Using the association of those data points, the technique builds the decision tree.

Step 3: Repeat step 1 and 2 after choosing how many trees you want to build.

Step 4: In case of the arrival of a new data point, make sure that each one of your trees predict the value for the data point, and assign the new point the total average across all of the predicted values.
4. Experimental Results

After selecting the four datasets, four classification techniques (Support Vector Machine, K-Nearest Neighbor, Random Forest, and Decision Tree) have been applied on those datasets to analyze which technique gives the best output (accuracy) on the same dataset. Each of this dataset has their own unique attributes such as the breast cancer dataset tries to detect whether the output is benign or malignant, the pima Indian diabetes tries to detect whether a particular person has diabetes or not based on some criterions, the heart dataset tries to find out if a particular person has heart problems on the basis of some attributes, the toddler asd dataset tries to evaluate whether a specific toddler has autism spectrum disorder traits based on some conditions. The whole experiment has been performed using Python Jupyter Notebook. As this kind of study has not been done before in the same approach, this experiment outputs unique result and provides better scope for the researchers and the students to experiment in this field to modify the result using different approaches.

The outputs (accuracy) of this experiment have been depicted below with graphs.

**BREAST CANCER DATASET**

![Graph showing accuracy of KNN, RF, SVM, and DT on Breast Cancer dataset.]

**Figure 1.** Applying KNN, RF, SVM, and DT on Breast Cancer dataset.

Figure 1 shows the accuracy of those given algorithms for the datasets of breast cancer; it indicates the accuracy of Random Forest is much better than the others considering two types (benign and malignant).
Figure 2. Applying KNN, RF, SVM, and DT on Pima Indian diabetes dataset.

Figure 2 shows the total accuracy of those given algorithms for the datasets of Diabetes; it indicates the accuracy of SVM is much better than the others considering a person has diabetes or not based on some criterions.

Figure 3. Applying KNN, RF, SVM, and DT on Heart dataset.

Figure 3 shows the calculated accuracy of those given algorithms for the datasets of Heart disease; it indicates the accuracy of Random Forest is much better than the others considering a person has any heart problem or not based on some conditions.
Figure 4 shows the accuracy of those given algorithms for the datasets of ASD; it indicates the calculated accuracy of Decision tree is much better than the others considering a particular child has autism spectrum disorder based on some criterions.

5. Conclusion
Now a days, people are falling sick very easily and they do not even know how or why or when they fall sick. So, by predicting or analyzing these different types of health-related datasets will make people more aware of their daily activities in their day to day life. Global community can be benefitted through this. In this study, we have analyzed performance of different types of techniques such as Random Forest, Support Vector Machine, K-Nearest Neighbor, and Decision Tree on some health-related datasets. But, along with this approach, we can also use lots of different types of algorithms and datasets to see whether they perform better than this study. This will help to realize the efficiency of performance of algorithms as well as will make people innovate new ideas to transform their lives in a healthy manner.

6. References
[1] Kaggle.com. (2019). Breast Cancer Wisconsin (Diagnostic) Data Set. [online] Available at: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data [Accessed 28 Nov. 2019].
[2] Kaggle.com. (2019). Pima Indians Diabetes Database. [online] Available at: https://www.kaggle.com/uciml/pima-indians-diabetes-database [Accessed 28 Nov. 2019].
[3] Kaggle.com. (2019). Heart Disease UCI. [online] Available at: https://www.kaggle.com/ronitf/heart-disease-uci [Accessed 28 Nov. 2019].
[4] Kaggle.com. (2019). Autism screening data for toddlers. [online] Available at: https://www.kaggle.com/fabdelja/autism-screening-for-toddlers [Accessed 28 Nov. 2019].
[5] Senturk, Z.K. and Kara, R., 2014. Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering, 4(1), p.35.
[6] Kala, R., Shukla, A. and Tiwari, R., 2009, December. Comparative analysis of intelligent hybrid systems for detection of PIMA indian diabetes. In 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC) (pp. 947-952). IEEE.
[7] Vijayarani, S. and Dhayanand, S., 2015. Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, 4(4), pp.816-820.

[8] Venkatalakshmi, B. and Shivsankar, M.V., 2014. Heart disease diagnosis using predictive data mining. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(3), pp.1873-7.

[9] Dhamodharan, S., 2014, May. Liver disease prediction using bayesian classification. In *4th National Conference on Advanced computing, applications & Technologies* (pp. 1-3).

[10] Lohita, K., Sree, A.A., Poojitha, D., Devi, T.R. and Umamakeswari, A., 2015. Performance analysis of various data mining techniques in the prediction of heart disease. *Indian Journal of Science and Technology*, 8(35), pp.1-7.

[11] Ramana, B.V., Babu, M.S.P. and Venkateswarlu, N.B., 2011. A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, 3(2), pp.101-114.

[12] Mandal, S.K., 2017. Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. *International Journal of Engineering and Computer Science*, 6(2), pp.20388-20391.

[13] Kahramanli, H. and Allahverdi, N., 2008. Design of a hybrid system for the diabetes and heart diseases. *Expert systems with applications*, 35(1-2), pp.82-89.

[14] Gupta, S., Kumar, D. and Sharma, A., 2011. Performance analysis of various data mining classification techniques on healthcare data. *International journal of computer science & Information Technology (IJCSIT)*, 3(4), pp.155-169.

[15] Chawarska, K., Macari, S.L., Volkmar, F.R., Kim, S.H. and Shic, F., 2014. ASD in infants and toddlers. *Handbook of Autism and Pervasive Developmental Disorders, Fourth Edition*.

[16] Abdullah, A.A., Rijal, S. and Dash, S.R., 2019, November. Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD). In *Journal of Physics: Conference Series* (Vol. 1372, No. 1, p. 012052). IOP Publishing.