International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

A Study to Evaluate Different Classifiers on the Basis of Performance of the Prediction for Major Common Diseases

Mohammad Munem Shahriar[1], Atiqul Islam Chowdhury[2]

[1]East West University, Department of Computer Science and Engineering, Dhaka, Bangladesh

[2]Ahsanullah University of Science & Technology, Department of Computer Science and Engineering, Dhaka, Bangladesh

Abstract: ***The world is changing day-by-day. But if there are no human to witness that change, then what the benefit is behind this change. As people are dying from some very common yet risky diseases, people need to be careful from not getting affected by any of those diseases. In this paper, we have worked on common diseases like Lung Cancer, Breast Cancer and Diabetes. They are some of those diseases that are affecting people every day and before preventing the outcome people are dying vigorously. As medical test can be time consuming, the data mining techniques can be very useful. In this study, we have implemented four classifiers on three kinds of datasets of the above mentioned diseases and predicted the performance. From those classifiers, Random Forest has given best results than the rest of the classifiers for the three diseases.***

Keywords: medical diseases, predictive data mining, classification, health data, cancer, diabetes

1. Introduction

There are so many serious diseases now-a-days. These diseases are tough to solve. Among the serious diseases, Brain cancer, Diabetes, Breast Cancer, Lung Cancer, Coronary Artery disease are most common. The Agency for Healthcare Research and Quality (AHRQ) clarifies that medical cost for cancer in the year 2011 in the United States was 88.7 billion dollars [1]. And out of different types of cancer, breast cancer has been one of the significant types over the past years [2]. The diabetes is another dangerous disease. It occurs when sugar levels in blood are too high. However, according to a study of Asian diabetic prevention organization, 60 percent of the whole world diabetic population is from Asia [3]. So, the risk is high for Asian people. An uncontrolled growth in tissues of the lung can be characterized as a malignant lung tumor which can cause lung cancer. It causes harm to the lung and it is one of the most dangerous diseases.

Medical decision could be extremely specialized and difficult job due to alternative factors or in case of rare diseases [4]. For deciding disease detection, the doctors have to test so many situations, issues for the test which is too much difficult. On the other side, if detection process is slow, then it may be harmful for the patient because some diseases have stages which have to be needed to determine. That's why our paper is about to predict the diseases like breast cancer, lung cancer and diabetes which are most common now-a-days. In this paper, we have manipulated medical data using various classification algorithms to optimize classifier performance for breast cancer, lung cancer and diabetes prediction.

The main objective of the research is to predict the different types of data with different values for cancers and diabetes. For the prediction process, we have to require nominal values of the dataset so that we can easily get the perfect result for the prediction. The prediction is based on some of the state of the art machine learning algorithms. The research has another objective that is to optimize the performances of these machine learning algorithms. That means the process that is Data Mining plays an important role to predict these diseases. The predictions and classifications in data mining help discover relations and patterns in patient medical data in order to improve their health [5]. Data Mining has two parts; predictive and descriptive. We have used predictive data mining which is used for classification and regression. The process is done by training data with different classifiers, analyzing the result and by accuracy visualization.

2. Related Study

Medical diagnosis system plays important role to determine the symptoms and disease type of a person, which involves classification test. There are so many researches on disease prediction, but important diseases were not focused on those papers. In this era, diabetes, breast cancer, lung cancer are the most common diseases and these diseases harm a lot to patients. We have studied different kind of research article based on disease prediction. A paper was about to detect symptom based diseases. Their system uses service oriented architecture (SOA) whereby the system elements of diagnosis, data portal and alternative miscellaneous services are provided [4]. The diseases were predicted by detecting the symptoms which are harmful for patient body. Another paper outlines the idea of predicting a particular disease by performing operations on the digital data generated in the medical diagnosis [5]. They used an efficient genetic algorithm hybrid with the techniques like back propagation and Naive Bayes approach for disease prediction is proposed. We have studied a paper which analyzed data mining techniques that could be used for predicting different types of diseases; which mainly concentrate on predicting heart disease, Diabetes and Breast cancer [6]. They used Naïve Bayes and Decision Tree (J48) algorithm to get the

accuracy. We have reviewed some papers which were about different diseases in different papers. Some paper focused heart disease, some focused diabetes etc. Some papers focused on multiple diseases with different classifiers. But with the changing of time, some diseases are cured in easy way. The disease condition and situation change with the flow of time. So we have focused mainly breast cancer, lung cancer and diabetes, which are now common and dangerous too. A paper used data mining framework which proposed two stages namely clustering and classification [7]. Cluster-0 and cluster-1, these are the two clusters that are generated by the First stage. There are no disease symptoms in Cluster-0 whereas cluster-1 has symptoms. Data analysis in medical system becomes broader now-a-days. We can classify the data by applying different types of algorithms to get the perfect result or accuracy. Our study is about to classify the cancer and diabetes data with various algorithms like Decision Tree, Support Vector Machine (SVM), Random Forest and Perceptron. Popular data mining algorithms (Support Vector Machine Analysis, Artificial Neural Network, Naive Bayes) are frequently used by the practitioners to develop a prediction model using attributes and attribute values [5]. Different papers use different algorithms. Some authors used just one algorithm on any particular dataset to predict the disease accuracy. We analyzed a paper in which Abraham proposed a methodology so as to increase classification accuracy of medical data based on Naive Bayes classifier algorithm [8]. So after reviewing those papers, we decided to work on different classification algorithms for different diseases which are dangerous and common in this era.

3. Methodology

In our proposed method, we have worked on three diseases. Those diseases are: Lung Cancer, Breast Cancer, and Diabetes. As all of these three diseases are different from each other, so we could not use only one dataset. Rather three different datasets of three diseases have been used here. We have wanted to predict each disease. For that reason, we have measured accuracy of the classification algorithms that have been used for each of the datasets. To measure the disease prediction accuracy of these datasets, four classification algorithms have been used. We have used the Weka tool to measure the accuracy [9]. Weka is a data mining tool which can measure the accuracy of the prediction rate of an algorithm running on a dataset [10]. A good visualization and clear interpretation of the attributes, features, instances and even the accuracy, F1 score, Recall, Precision can be displayed from Weka. It gives a complete visualization of what kind of data the dataset contains. This has helped us to understand the attribute and instances of the datasets more transparently. Fig. 1 shows the steps of our proposed work.

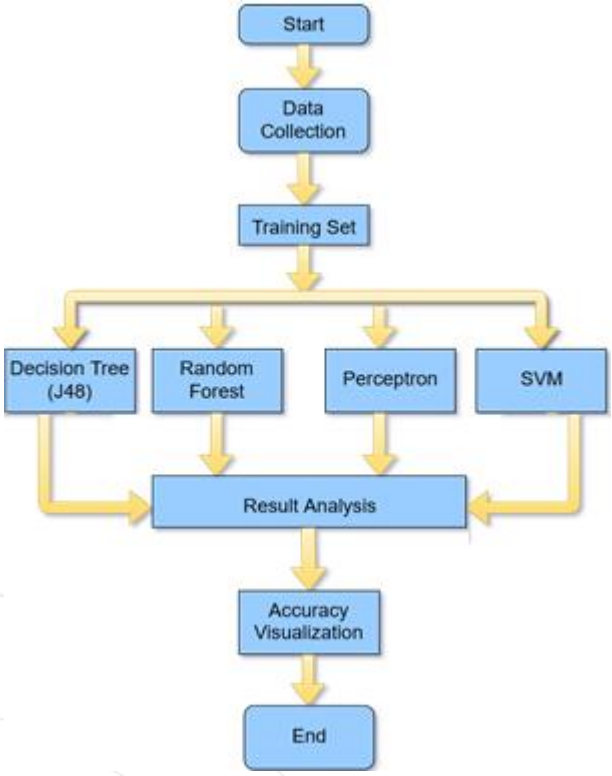


Figure 1: Flowchart of proposed method

For our approach, at first we collect the data. The data needs to be cleaned. Some datasets have missing values, some do not. Missing value can give erroneous result which can hamper the accuracy result. After that, when the data is cleaned, we have trained it to detect each disease. We have used 10-fold cross validation to train the datasets. Each time we have trained the data with different classification algorithm. After that we analyzed the result thoroughly of the accuracies. Then we compare each of the accuracy of the prediction result.

After visualizing the accuracy result, we have come to understand which algorithm is better for which disease for prediction.

4. Experimental Analysis

For our research, we have used four different classifier algorithms to get the accuracy of the diseases.

a) Data

Three kinds of datasets have been used here. We have used these datasets from a website that is very useful to use this resource [11]. The dataset for Lung Cancer consists of 16 attributes and 309 instances.

Table I: Elements of Lung Cancer Dataset

Attributes	Weight
Gender	M or F
Age	21 to 87
Smoking	1(yes) or 2 (no)
Yellow_Fingers	1(yes) or 2 (no)
Anxiety	1(yes) or 2 (no)
Peer_Pressure	1(yes) or 2 (no)
Chronic Disease	1(yes) or 2 (no)

Fatigue	1(yes) or 2 (no)
Allergy	1(yes) or 2 (no)
Wheezing	1(yes) or 2 (no)
Alcohol Consuming	1(yes) or 2 (no)
Coughing	1(yes) or 2 (no)
Shortness Of Breath	1(yes) or 2 (no)
Swallowing Difficulty	1(yes) or 2 (no)
Chest Pain	1(yes) or 2 (no)
Lung_Cancer	Yes or No

The dataset for Breast Cancer consists of 10 attributes and 286 instances.

Table II: Elements of Breast Cancer Dataset

Attributes	Weight
age	40 to 79
menopause	premeno, ge40, lt40
tumor-size	0 to 54
inv-nodes	0-2 or 15-17
node-caps	yes or no
deg-malig	1, 2 or 3
breast	right or left
breast-quad	left_up, left_low, right_up, right_low, central
irradiat	yes or no
Class	recurrence-events or no-recurrence-events

The dataset for Diabetes consists of 9 attributes and 768 instances.

Table III: Elements of Diabetes Dataset

Attributes	Weight
preg	0 to 17
plas	0 to 199
pres	0 to 122
skin	0 to 99
insu	0 to 846
mass	0 to 67.1
pedi	0.078 to 2.42
age	21 to 81
class	tested_positive or tested_negative

b) Classifiers Used

There are many classifier algorithms by which we can get the better accuracy on different dataset. We have mainly focused on:

- Decision Tree Classifier
- Random Forest
- Perceptron
- Support Vector Machine (SVM)

And TABLE IV lists the classifiers with the Weka built in name.

Table IV: Classifiers Used in Weka

Classifier	Weka Built in Name
Decision Tree Classifier	weka.classifiers.trees.J48
Random Forest	weka.classifiers.trees.RandomForest
Perceptron	weka.classifiers.functions.multilayerPerceptron
SVM	weka.classifiers.functions.supportVector

c) Result Analysis

We will show the experimental results in this part. First of all, we have used Decision Tree Classifier algorithm for different diseases. We have calculated accuracy, weighted average Precision, weighted average Re-call and weighted average F-measure for these four types of algorithms. TABLE V shows the result of Decision Tree Classifier for breast cancer, lung cancer and diabetes data.

Table V: Decision Tree Results

Measures	Lung Cancer Data	Breast Cancer Data	Diabetes Data
Accuracy	95.4693%	78.3217%	84.1146%
Weighted average Precision	0.953	0.779	0.842
Weighted average Re-call	0.955	0.783	0.841
Weighted average F-measure	0.954	0.760	0.836

After that, Random Forest is applied for these different data. We have got 100% accuracy for the diabetes data. The result is shown in TABLE VI.

Table VI: Random Forest Results

Measures	Lung Cancer Data	Breast Cancer Data	Diabetes Data
Accuracy	99.6764%	97.9021%	100%
Weighted average Precision	0.997	0.979	1.00
Weighted average Re-call	0.997	0.979	1.00
Weighted average F-measure	0.997	0.979	1.00

Next, we have applied Perceptron algorithm for getting the result. The accuracy is quite good for Lung Cancer and Breast Cancer data, but not so good for Diabetes data. TABLE VII shows the result for Perceptron algorithm.

Table VII: Perceptron Results

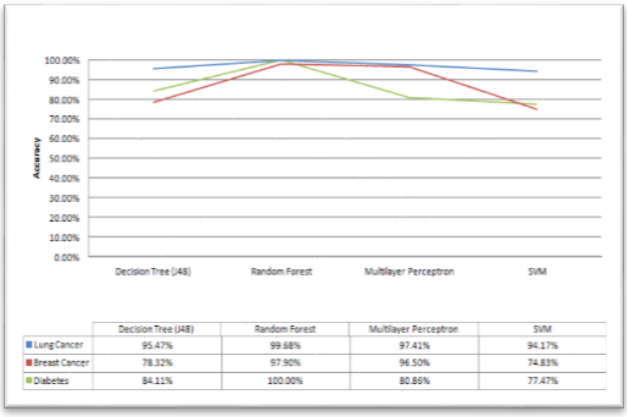
Measures	Lung Cancer Data	Breast Cancer Data	Diabetes Data
Accuracy	97.411%	96.5035%	80.8594%
Weighted average Precision	0.974	0.965	0.821
Weighted average Re-call	0.974	0.965	0.809
Weighted average F-measure	0.974	0.965	0.812

After experimenting these three classifier algorithms, we have decided to apply SVM on these three datasets. But the result is down for this classifier. The result is shown in TABLE VIII.

Table VIII: Support Vector Machine (SVM) Results

Measures	Lung Cancer Data	Breast Cancer Data	Diabetes Data
Accuracy	94.1748%	74.8252%	77.474%
Weighted average Precision	0.940	0.731	0.771
Weighted average Re-call	0.942	0.748	0.775
Weighted average F-measure	0.940	0.723	0.764

From these four algorithms, we have got better accuracy for Random Forest algorithm. And this will get clear if we see the plot of these algorithms (shown in Fig.2) applied on different datasets.



	Decision Tree (J48)	Random Forest	Multilayer Perceptron	SVM
Lung Cancer	95.47%	99.68%	97.41%	94.17%
Breast Cancer	78.32%	97.90%	96.50%	74.83%
Diabetes	84.11%	100.00%	80.86%	77.47%

Figure 2: Accuracy Comparison

Lastly, we can say that we have got better accuracy for Random Forest algorithm among those four classifier algorithms. So prediction using this Random Forest algorithm will give better accuracy and result for predicting those diseases.

5. Future Work

We have performed prediction with the help of four classifiers. From which for diabetes, Random Forest has given 100% accuracy. To determine if there are any classifiers which can give more accurate result for the other two algorithms, we want to use some more classifiers other than those four classifiers onto these datasets to observe the accuracy. We also want to use these algorithms onto other large datasets for future to observe whether those classifiers can give same amount of accuracy result on large datasets [12]. Further that, we have used only classifiers. We will try some other data mining technique like Clustering, Time Series, and Assertion Rules on the datasets to observe the result.

6. Conclusion

In this study, we have brought light to the diseases which often do not bother us, but they surely take millions of lives each year. Medical science is developing each day. So the methods to detect and predict diseases should also be improved, thus comes data mining techniques for prediction [13]. We have found best result for diabetes by using random forest. There are some chronic diseases which also needs prediction method to prevent the occurrence of these diseases. These data mining techniques can ease the work of the doctors and medical scientists to detect and predict any diseases in a very short period of time. These techniques are needed more for the future for any upcoming epidemic events.

References

[1] American Cancer Society, “Cancer Facts & Figures 2015,” Cancer Facts Fig. 2015, pp. 1–9, 2015.
[2] BCSC, “Types of Breast Cancer,” Breast Cancer Society of Canada, 2014. Available: http://www.bcsc.ca/p/41/l/506/t/Breast-Cancer-Society-of-Canada---Types-of-Breast-Cancer.
[3] J. C. N. Chan, V. Malik, W. Jia, T. Kadowaki, C. S. Yajnik, K.-H. Yoon,and F. B. Hu, “Diabetes in Asia: epidemiology, risk factors, and pathophysiology.,” JAMA, vol. 301, no. 20, pp. 2129–40, 2009.
[4] B. S. Sathyabama Balasubramanian, "SYMPTOM’S BASED DISEASES PREDICTION IN MEDICAL SYSTEM," *International Journal of Advances in Computer Science and Technology*, vol. 3, no. 2, pp. 123-128, 2014.
[5] H. S. N. F. S. Ajinkya Kunjir, "A Review on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques," *International Journal of Computer Applications*, vol. 155, no. 1, pp. 34-38, 2016.
[6] D. D. S. P. K.Gomathi, "Multi Disease Prediction using Data Mining," *Online. Available: http://www.publishingindia.com*, 2016.
[7] B.M.Patil, Ramesh C.Joshi and Durga Toshniwal, "Effective framework for Prediction of Diseases outcome using medical Datasets clustering and classification," *Published in International Journal of computational Intelligence* studies, Vol 1, Issue 3, pages 273-290, August 2010.
[8] Ranjit Abraham, Jay B.Simha, Iyengar, “A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier”, In proc of IEEE international conference on information Technology, pp. 235 - 236, 2006, ISBN: 0-7695-2635-7.
[9] Ajinkya kunjir, Harshal sawant, Nuzhat Sheikh, “Data mining and visualization for prediction of multiple diseases in healthcare”, IEEE 2007.
[10] Munaza Ramzan, “Comparing and Evaluating the Performance of WEKA Classifiers on Critical Diseases”, 978-1-4673-6984-8/16/$31.00 2016 IEEE.
[11] Datasets are collected from here: https://.data.world/
[12] Weimin Xue, Yanan Sun, Yuchang Lu, “Research and Application of Data Mining in Traditional Chinese Medical Clinic Diagnosis”, In proc of IEEE 8th international Conference on Signal Processing, Vol. 4, ISBN: 0-7803-9736-3, 2006.
[13] Boris Milovic and Milan Milovic, Prediction and Decision Making in Health Care using Data Mining, IJPHS, 2012, 69-78.