Predictive Analytical Approach for Disease Detection using Machine Learning

Ayman Mir[1] and Prof A. A Godbole[1]

Sardar Patel Institute of Technology, Department of Computer Engineering, Mumbai, India
ayman.mir@spit.ac.in
anand_godbole@spit.ac.in

Abstract. Rapid technological advancement is increasing in all domains including healthcare. Machine learning in healthcare is growing tremendously. Automated disease diagnosis using predictive analytics has set foot in healthcare. There is a need to figure out the best performing algorithm using predictive analytics and to test it on various disease dataset. The diseases that would be taken into consideration is Diabetes, Chronic Kidney Disease, Heart Disease. Few of popular machine learning algorithms are taken into consideration. The proposed approach will result into comprative analysis for best performing algorithm.

Keywords: Machine Learning· Naive Bayes· Support Vector Machine· Random Forest· Simple CART · KNN

1 Introduction

Currently in the health care domain machine learning is being used to find meaningful patterns from the huge data gathered in terms of disease dataset, medical reports etc. Healthcare domain provides a lot of scope for research as it has tremendously evolved. Machine learning approach can be applied for prediction of diseases and provide automated diagnosis under the validation of professional doctor. Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in efficient diagnosis .

Machine learning will help analyze the huge healthcare data and find insightful patterns also machine learning can aid in reducing the ever increasing cost of healthcare. Machine learning can be applied to various sub domains of healthcare which will help doctors determine more informed and personalized prescriptions and treatments for patients. Through the usage of machine learning in medical domain of healthcare it will help the physicians to make informed and accurate diagnoses.

The remainder of the paper is organized as follows :

- Section 2 is Literature Survey describing the already existing work.
- Section 3 is Methodology describing the workflow of proposed methodology.

- Section 4 describes the Experimental results that are obtained after building classifiers through WEKA and discusses the performance evaluation .
- Section 5 is about Conclusion which concludes the overall results.

2 Literature Survey

This section reviews the existing recent literature work and provides insights in understanding the challenges and tries to find the gaps in existing approaches.

Following is the comparison of various methodology along with outcome of the various existing research work

Table 1. Comparison of various Literature Works

Ref	**Methodology**	**Dataset**	**Outcomes**	**Limitations**
Bhargava et al. 2017 [1]	Simple CART in WEKA to predict heart attack	Real world Male Heart disease dataset. Instances Used =209	Accuracy of correctly classified instances is 79.9 %	Only one algorithm used hence no comparisons for better accuracy.
Dhomse, Mahale, 2016 [2]	SVM, Decision Tree and Naive Bayes used to predict heart disease using WEKA	Heart disease dataset from Cleveland Clinic Foundation. Instances Used =303	After reducing dataset SVM outperforms Naive Bayes	Accuracy results not mentioned direct graph plotted.
Dhomse, Mahale, 2016 [2]	SVM and Naive Bayes used to predict diabetes disease using WEKA tool	Diabetic patients dataset is collected from hospital repository. Instances Used = 1865	Naive Bayes has better accuracy along with reduced time for building the training model than SVM	Accuracy of Naive Bayes is 34.89% which is quite risky for prediction
Ramzan, 2016 [3]	Naive Bayes, J48 Decision Tree, Random Forest are used to compare classifiers to predict critical disease using WEKA.	Disease classification dataset collected from Global Health Data Exchange. Instances Used = 9242	Random Forest turns out with an Accuracy of 99.83% beating both Naive Bayes and J48	Random Forest requires more time for building the training model
Naik, Samant, 2016 [4]	Decision Tree, KNN, Naive Bayes are used to predict liver disorder using WEKA, Orange, Tungara, KNIME and Rapid miner tool	Liver patient dataset collected from Indian Liver Patient Dataset. Instances Used = 583	KNIME tool's performance was the best. Decision Tree and KNN outperformed Naive Bayes using all the tools	Requires a powerful machine learning to analyze the outcome of model using all tools to improve classification accuracy.
Iyer et al., 2015 [5]	J48 Decision Tree and Naive Bayes approach for diagnosis of diabetes	Pima Indians Diabetes Database. Instances Used = 768	Naive Bayes gives least error rate and thus outperforms J48.	Comparison of only 2 algorithm is not sufficient.

Rohan Bhardwaj et al. [6] summarizes the potential changes that machine learning can bring about in healthcare sector. Though it will not replace the physician it will bring about positive transformation in the current sector of healthcare. By implementing machine learning interesting patterns in disease data are discovered. It highlights the significance of machine learning in healthcare sector.

Niharika G. Maity et al. [7] suggested for improved diagnosis and prognosis machine learning can be used effectively and demonstrated with two case studies for disease diagnosis of Alzheimer's and Classification of Cancer.

Dhafar Hamed Abd et al. [8] applies machine learning approach to E-medication system for Sickle Cell Disease diagnosis where it tries to bridge the gap between the doctor and patient through application on the smart phone.

Parisa Naraei et al. [9] gives comparison analysis performed using SVM and MLP neural networks for heart disease prediction where SVM gives higher accuracy. SVM outperformed other machine learning algorithms with the highest accuracy among all. Since there are various machine learning techniques the research makes comparison so as to determine which technique gives efficient results.

3 Methodology

This section includes the methodology describing the approach that is used to carry out the research in order to perform comparative analysis

3.1 Flow Chart of Proposed Methodology

The proposed methodology is different from the already reviewed work in terms of considering multiple disease datasets along with various supervised and unsupervised algorithms to test the performance and evaluate the results.

The Proposed Predictive Analytical model considers three disease datasets as input for Heart, Diabetes and Chronic Kidney disease. The input dataset is processed using popular machine learning algorithms that are Naive Bayes, SVM, Random Forest, Simple CART and KNN for each algorithm respective classifier model is trained and tested. Based on the experimental results the best performing algorithm can be determined. The following describes the steps involved in the procedure of the Fig 1. Proposed Classifier Methodology

Step wise Procedure of Proposed Methodology

- **Step 1 :** Preprocess the input datasets for all three diseases.
- **Step 2 :** Divide dataset of respective diseases into Training set and Test set based on cross validation or percentage split.
- **Step 3 :** Select the machine learning algorithm i.e. Naive Bayes, Support Vector Machine, Random Forest, Simple CART and KNN algorithm.
- **Step 4 :** Build the predictive analytical model for each mentioned machine learning algorithm based on training set.

- **Step 5 :** Test the Classifier model for the mentioned machine learning algorithm based on test set
- **Step 6 :** Perform Comparison Evaluation of the experimental performance results obtained for each predictive model.
- **Step 7 :** After analyzing based on various measure conclude the best performing algorithm.

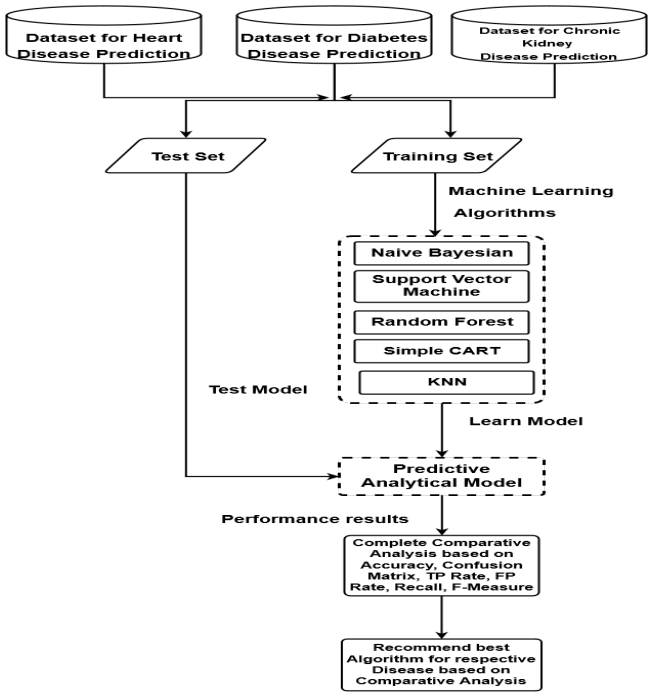

Fig. 1. Proposed Methodology Flowchart

The proposed classifier model has been built using WEKA tool and based on successful execution of each step we can evaluate the experimental results.

3.2 Datasets Used

Here is the description of the datasets that will be used as an input to predictive analytical model

1. **Diabetes Dataset:** "Pima Indians Diabetes Database", Instances : 768, Attributes: 9
2. **Heart Dataset:** "UCI Heart Disease Dataset", Instances : 270, Attributes: 14
3. **Chronic Kidney Dataset:** "UCI CKD Disease Dataset", Instances : 400, Attributes: 25

4 Experimental Results

This section describes the experimental results that are obtained after training the predictive analytical model.

According to the Experimental Results for Heart Disease the Accuracy of Naive Bayes algorithm is the highest which is 85.18% and next is SVM having accuracy 82.40% whereas Random Forest and Simple CART algorithm have equal accuracy of 79.62% and KNN has 72.23%

According to the Experimental Results for Diabetes Disease the Accuracy of SVM is the highest which is 78.50% and Naive Bayes and Random Forest have almost equal accuracy of 75% whereas Simple CART has an accuracy of 74.26% and KNN has an accuracy of 69.7068%

According to the Experimental Results for Chronic Kidney Disease the Accuracy of Random Forest is the highest which is 100% and Simple CART algorithm has the second highest accuracy of 98.12%. The accuracy of SVM is 96.87% and the accuracy of Naive Bayes is 95% and KNN is 94.37% for the classification of chronic Kidney Disease.

Following is the graph in Fig 2 demonstrating the accuracy measure value of the best performing algorithm of the respective disease.

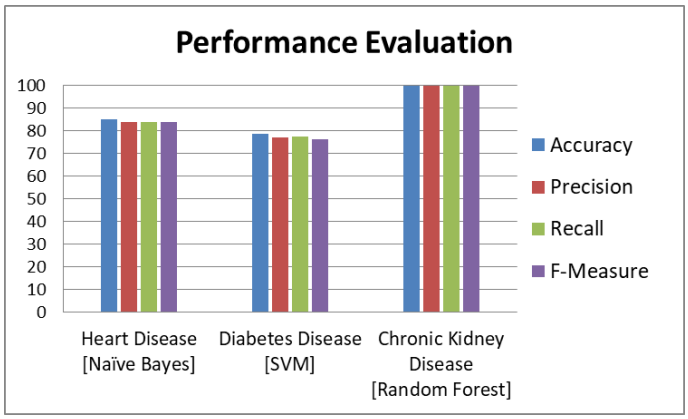

Fig. 2. Performance Evaluation of the Proposed Methodology

It can be observed that for Heart Disease dataset the best performing algorithm is Naive Bayes. For Diabetes disease dataset it is SVM and for Chronic Kidney Disease it is Random Forest.

Also it can be observed that throughout the predictive analytical model for respective diseases the performance of each algorithm varies. It can be observed that SVM and Random Forest has performed well with each disease datasets.

5 Conclusion

This section concludes the research intended and summarizes the approach.

In this research work for building a predictive analytical model popular machine learning algorithms such as Naive Bayes, SVM, Random Forest, Simple CART and KNN has been applied using WEKA. In this research multiple disease datasets has been taken into consideration which are Heart Disease, Diabetes Disease, Chronic Kidney Disease. The performance of each algorithm is tested based on the classification accuracy along with other performance evaluation parameters Precision, Recall and F-Measure.

It is observed as per Performace Evaluation that for Heart Disease dataset Naive Bayes outperformed others. For the Diabetes disease dataset the SVM outperformed the others. For the Chronic Kidney Disease the Random Forest outperformed the others.

Also the research can be extended by using more cleaned and unbiased disease datasets from different authentic sources and testing based on real time data. Also various other diseases can be taken into consideration and tested.

References

1. N. Bhargava, S. Dayma, A. Kumar and P. Singh, "An approach for classification using simple CART algorithm in WEKA", 2017 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2017, pp. 212-216.
2. Dhomse Kanchan B. and Mahale Kishor M., "Study of machine learning algorithms for special disease prediction using principal of component analysis," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, 2016, pp. 5-10.
3. M. Ramzan, "Comparing and evaluating the performance of WEKA classifiers on critical diseases," 2016 1st India International Conference on Information Processing (IICIP), Delhi, 2016, pp. 1-4.
4. Amrita Naik, Lilavati Samant, Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime, Procedia Computer Science, Volume 85, 2016, Pages 662-668, ISSN 1877-0509
5. Iyer, Aiswarya Jeyalatha, S Sumbaly, Ronak. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining Knowledge Management Process. 5. 1-14.
6. R. Bhardwaj, A. R. Nambiar and D. Dutta, "A Study of Machine Learning in Healthcare," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, 2017, pp. 236-241.
7. N. G. Maity and S. Das, "Machine learning for improved diagnosis and prognosis in healthcare," 2017 IEEE Aerospace Conference, Big Sky, MT, 2017, pp. 1-9.
8. D. Abd, J. K. Alwan, M. Ibrahim and M. B. Naeem, "The utilisation of machine learning approaches for medical data classification and personal care system mangementfor sickle cell disease," 2017 Annual Conference on New Trends in Information Communications Technology Applications (NTICT), Baghdad, 2017, pp. 213-218.
9. P. Naraei, A. Abhari and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," 2016 Future Technologies Conference (FTC), San Francisco, CA, 2016, pp. 848-852.