International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Comparison of Machine Learning Algorithms for Detection of Diabetic Retinopathy with Hybrid Feature Selection

Amalu Michael[1], Deepa S S[2]

P.G. Student, Department of Computer Science and Engineering, Govt: Engineering College, Idukki, Kerala, India[1]

Associate Professor, Department of Computer Science and Engineering, Govt: Engineering College, Idukki, Kerala, India[2]

ABSTRACT:Diabetic Retinopathy (DR) is an eye disease that occurs due to a large amount of glucose level in blood which forms damages in retinal vessels. Early detection of diabetes is an efficient way to prevent DR condition. Various methods are used for detecting DR. Medical imaging is one of them but it is more expensive and requires advanced equipment. Detection of DR through low-cost, easy available electronic health records is more convenient. In this paper, we propose a diabetic retinopathy detection model using EHR with a hybrid feature selection method. Feature selection is a process of finding the most important feature subset for the ML application. Here a hybrid feature selection that combines both correlation and PSO is used. Then the selected features are applied to five classifiers which are NB, LR, DT, RF, and SVM to compare their performance. From the experimental results, it is found that the RF classifier achieves an accuracy of 83% with a reduced set of features.

KEYWORDS:Machine Learning, Diabetic Retinopathy, Feature Selection, Electronic Health Record.

I. INTRODUCTION

Diabetic retinopathy (DR) is the damage of blood vessels in the retina due to poorly controlled blood sugar. Nowadays, DR disease becomes the common cause of blindness. The main symptoms of diabetic retinopathy are difficulty in color vision, total vision loss and blurred vision etc. There are two main type of DR disease which are non-proliferative (NPDR) and proliferative (PDR). The non-proliferative is generally symptoms-free and proliferative is the advanced stage which usually leads to blindness. Regular analysis of blood sugar level of patient is an effective method to prevent the chance of Retinopathy disease.

An Electronic Health Record (EHR) is a medical information collection which stored in digital format and used in various disease detection and prediction applications. The advantages of using EHR is that it is very easy to get at low cost. So regular screening of diabetic retinopathy through EHR is more convenient than other methods.

Machine Learning (ML) has an efficient and effective disease diagnosis system using a set of features. The algorithms of Machine Learning are useful for finding patterns from large amount of data. This facility is well suited for various field especially for medical applications. Machine Learning is a part of artificial intelligence that provide the ability to automatically learn from experiences and it mainly focuses on the creation of programs that can access data and use this data to create patterns. The aim of machine learning is to allow the computer to automatically learn and create patterns without human interactions and make decision about future using these patterns.

Feature selection is a machine learning procedure to select a feature subset that are most important for the machine learning application. Elimination of irrelevant features reduces the complexity of model and training time of classifiers because the number of irrelevant features are proportional to the training time of classifier. The main objective of feature selection is to improve the classification performance with minimum number of features. Feature selection process is generally classified into four, which are filter, wrapper, embedded and hybrid. Filter approach find relevance of features by some calculated values such as correlation, standard deviation etc. The wrapper approaches are based on algorithms which evaluate feature subsets and selects best one. Embedded approach finds features based on classifiers. Hybrid approach is the combination of both filter and wrapper approach.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

This paper is organized as follows. Section II deals with summarizing the related work done under the scope of this paper. The section III deals with the overall design of the proposed method. Section IV briefs the analysis of result of the proposed method. Section V conclusion

II. RELATED WORK

Diabetic Retinopathy detection is one of the main task in medical field which is mainly based on various image processing techniques. G G Gardner et al,[10] presents an automated DR diagnosis system using image processing techniques. In this work a back propagation neural network is used to analyses the features of retinal fundus images. Adarsh et, al [12] presents a DR detection system based on image processing and SVM classifiers. Here, the features are extracted by image processing techniques and these are applied to SVM classifier to classify them as infected or uninfected. Rubya Afrin et, al [9] uses image processing and fuzzy logic to detect DR from retinal images. Here three lesion features and two textural features are extracted from fundus images. Fuzzy logic is used to find the exact stages of DR disease. Carrera et, al [15] presents an automated retinopathy system using image processing to extract features and SVM to find the presence of retinopathy. Kranthi Kumar et, al [18] introduces an automatic method for DR detection using digital image processing. Here the DR is detected by the presence of exudate. For exudate detection a combination of back ground subtraction with exudate candidate extractions is used. Here an accuracy of 82% is obtained. R.Priya et, al [17] proposes a DR detection system based on Image processing and neural networks. The features like blood vessels, haemmoraghes finds the amount of disease spared in retina. Here these features are extracted by boundary detection and morphological processing and the extracted features are applied to classifiers for classifying the disease stage.

DR detection is mainly based on images. Nowadays various disease detection using Electronic Health Record is more popular. Zheng et, al [7] proposes type 2 diabetics detection system based on EHR. Here, the most important features of EHR is selected via feature engineering and the selected features are applied to six machine learning algorithms to evaluate their performance. Among the six algorithms the RF algorithm provides more accurate results as compared with others. N Sneha et, al [6] proposes a Diabetics Mellitus prediction system using clinical records. Here good features for this application are selected based on correlation value between features. [8] is a review paper about EHR and its applications in research and clinical applications. The advantages of using EHR is that these are very easy to obtain and less expensive. Yunlei Sunet, al [3] proposes a diabetic retinopathy detection system based on EHR. This method is more convenient and cost effective as compared with medical imaging. Five machine learning classification algorithms are used. Among that the RF classifier achieves highest accuracy.

DR detection using Electronic Health record is more convenient than medical imaging. Feature Selection improves the performance and reduces the overhead of machine learning algorithms. Because the presence of irrelevant features increases the complexity of model and the training time of classifiers. So it is necessary to selecting most important features from EHR data.

III. SYSTEM DESIGN

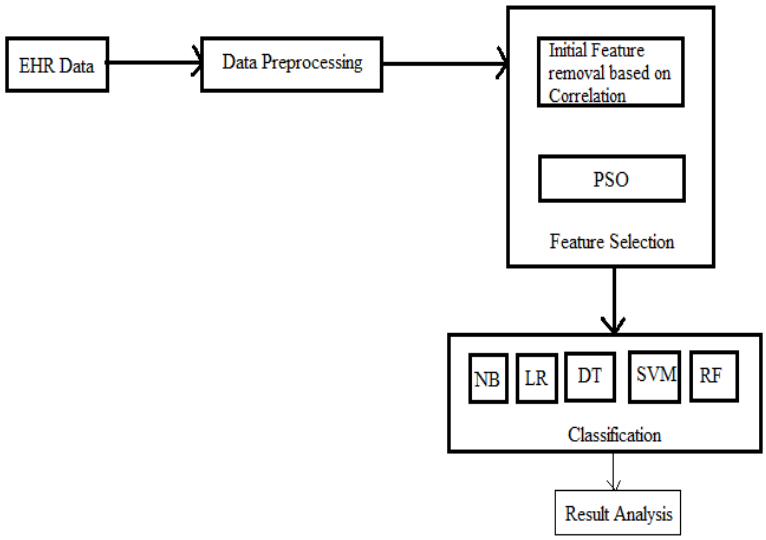
The proposed diabetic retinopathy detection system includes four modules which are Data preprocessing module, Feature selection module and Classification module as shown in fig 1.

(i)Data Preprocessing

Data preprocessing is actually the process of cleaning data. This work is implemented by using real world Electronic Health Record (EHR). The EHR data is the systematic collection of patient's medical information which consist of test results and other health related data. The real world data may be noisy, incomplete and inconsistent. For better diagnosis these data need to be handled. During these stage the features with more than 45% is removed otherwise the missing values are replaced by mean value of that feature. It also checks duplicate records and text data values.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||



(ii). Feature Selection:

A hybrid feature selection which combines correlation and Particle Swarm Optimization is used. First step (Correlation) is a filter approach of feature selection. Correlation shows how the features are related to the target variable. Here the correlation value of each feature with class is computed then the features are ranked based on this value. The features which have correlation value below some threshold value are eliminated.

In the second step, the most informative features from the result of first step are selected using PSO. PSO is actually an optimization algorithms which iteratively trying to find best solution to a problem. One of the drawback of PSO is that it has high computational cost and it does not consider the feature redundancy while generating feature subsets for evaluation. So some pre- filtering process is very necessary for PSO. Here the feature subsets are considered as the particles. In each iteration each subset is applied to LR classifier to evaluate its accuracy and also the cost of execution in calculated. Finally the subset with least cost and high accuracy is selected as the optimal subset of features.

(ii). Classification:

In this work five widely used classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are used. These classifiers are used to evaluate the performance of feature selection. Here a DR Dataset of 12 features and 2050 records are used. The data is divided in 80:20 ratio as train and test data respectively.

IV. EXPERIMENTAL RESULTS

A. *Diabetic Retinopathy Dataset*

The dataset is downloaded from the kaggle site. Dataset consist of 12 attributes. Theyare Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetic Pedigree Function, Triceps skinfold thickness, BOI, Hr2 serum insulin, Plasma glucose concentration 2 hr, Age and the label. It contains 2050 rows and 12 columns. The data is divided in 80:20 ratio as train and test data respectively. The train data is used identify the relationship between each feature and class label to create model for disease detection. The test data is used to evaluate the performance of model crated.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

No	Attribute
1	Pregnancies
2	Glucose
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	BMI
7	Diabetic Pedigree Function
8	Triceps skinfold thickness
9	BOI
10	Hr2 serum insulin
11	Plasma glucose concentration 2 hr
12	age

B.Result Analysis

The main aim of this work is to create a machine learning model for the diagnosis of diabetic retinopathy based of EHR data with hybrid feature selection. To evaluate the performance of this hybrid feature selection based of correlation and PSO, a DR data set with 12 features were used. For the experimental analysis, the data without any feature selection is applied to five classifiers. Without feature selection Support Vector Machine and Decision Tree provides better accuracy than the other three classification models, both of these models provide an accuracy of 81%. The same data after feature selection is also applied to these classification models to compare their results. After the data is processed for feature selection, the SVM still keeps a high accuracy. By comparing the data before and after the implementation of the hybrid feature selection method, it is found that the diagnostic accuracy of Random Forest (RF) and Logistic Regression (LR) has been improved after feature selection as shown in table 1 and figure 2.

Classifier	Without Feature Selection	With Feature Selection
RF	78.09	82.09
SVM	81.50	81.51
LR	76.87	81.46
DT	81.06	78.45
NB	79.31	79.31

Table 1: Classification accuracy before and after feature selection

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

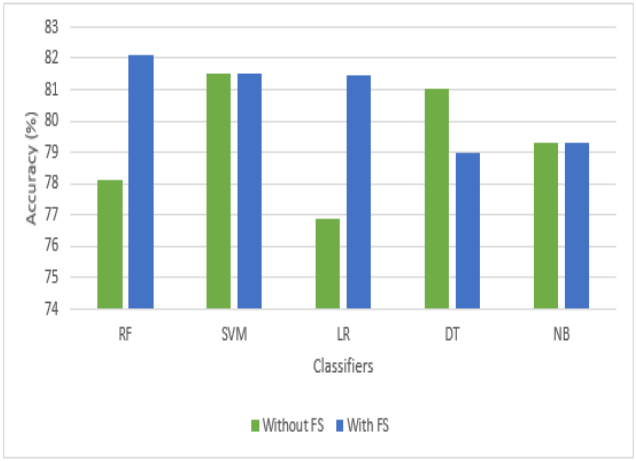


Figure 2: Classification accuracy

The accuracy of RF is improved from 78% to 82%, while the accuracy of LR is improved from 76% to 81%. However, the diagnostic accuracy of SVM is not affected by the characteristics of this feature selection. NB classification model provides the same classification accuracy before and after feature selection. After performing feature selection, DT reduced its accuracy from 81% to 78%. Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

Classifier	Filter approach (Correlation)	Hybrid approach (Corre+PSO)
RF	81.75	82.09
SVM	81.5	81.51
LR	78.29	81.46
DT	81.8	78.45
NB	79.17	79.31

Table 2: Comparison of filter and wrapper approach

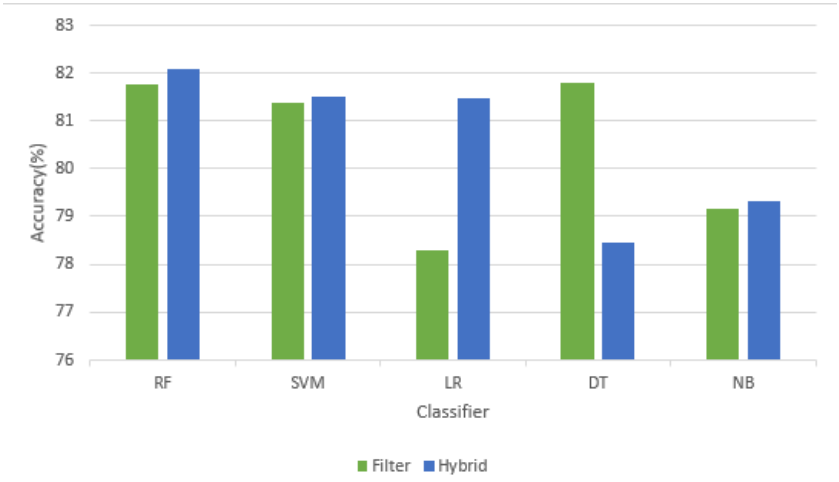


Figure 3: Comparison of classification accuracy between filter and hybrid approach

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Table 2 and figure 3 shows the advantage of the hybrid algorithm over the filter approach. The accuracy of the LR algorithm is improved from 78% to 81%, and the RF algorithm from 81% to 82% by reducing features from 10 to 6.

V. CONCLUSION

Diabetic retinopathy is an eye condition that occurs due to poorly controlled blood sugar level. It causes damage to the blood vessels in retina and nowadays it becomes the common cause of blindness.Early diagnosis of diabetes by Electronic Health Records (EHR), has become an effective way to prevent Diabetic Retinopathy disease. The algorithms of Machine learning are helpful for identifying complicated patterns within prosperous and huge data. These patterns are used in numerous illness diagnosis and prediction. In this work, a Diabetic Retinopathy detection model using Electronic Health Record with hybrid feature selection method is completed. The DR detection using EHR data is more convenient and low-cost as compared with traditional DR diagnostic methods (via human fundus images). Among 12 features 6 features are removed by this hybrid feature selection. From the analysis of the experimental results, it is found that the most suitable machine learning model for DR disease diagnosis using EHR data is RF, LR and SVM. The RF classifier achieves an accuracy of 82%, both SVM and LR achieves an accuracy of 81%.

REFERENCES

[1] K. Shailaja et, al, “Machine Learning in Healthcare: Review” 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018)
[2] Raid Alzubi et, al “Hypertension Disease Detection via Machine Learning Techniques “IEEE Transactions on Neural Networks and Learning Systems 2018
[3] Yunlei Sun et, al “Diagnosis And Analysis Of Diabetic Retinopathy Based On Electronic Health Records” Special Section On Healthcare Information Technology For The Extreme And Remote Environments 2019
[4] Syed Sifat Rahman et, al “A Machine Learning-Based Approach for Diabetes Detection and Care” 2nd International Journal on Next Generation Computing Technologies (NGCT-2016)
[5] Yvan Saeys et, al “A review of feature selection techniques in bioinformatics” International Conference on bioinformatics 2016
[6] N. Sneha et, al “Analysis of diabetes mellitus for early prediction” Journal of Big data 2019
[7] Tao Zheng et, al “A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records” International Journal of Medical Informatics 2016
[8] Peter B. Jensen et, al “Mining electronic health records: towards better research applications and clinical care” 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018)
[9] R. Afrin and P. C. Shill, ``Automatic lesions detection and classification of diabetic retinopathy using fuzzy logic," in Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST), Jan. 2019, pp. 527_532. doi: 10.1109/ICREST.2019.8644123.
[10] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, ``Automatic detection of diabetic retinopathy using an artificial neural network:A screening tool," Brit. J. Ophthalmology, vol. 80, no. 11, pp. 940_944,1996.
[11] Sakshi Gujral “Early Diabetes Detection using Machine Learning” International Journal for Innovative Research in Science & Technology 2017
[12] P. Adarsh and D. Jeyakumari, ``Multiclass SVM-based automated diagnosis of diabetic retinopathy," in Proc. Int. Conf. Commun. Signal Process. (ICCSP), Apr. 2013, pp. 206_210.
[13] Swati Gupta et, al “Diagnosis of Diabetic Retinopathy using Machine Learning” Journal of Research and Development 2015
[14] Karan Bhatia et, al “Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm” International Conference on Next Generation Computing Technologies (NGCT-2016)
[15] Enrique V. Carrera et, al “Automated detection of diabetic retinopathy using SVM” International Journal on Computer, Communication, Control 2017
[16] Dhiravidachelvi, E. et, al “Diagnosing Diabetic Retinopathy in Fundus Images” Journal of Computer Science 2018
[17] R. Priya et, al “Diagnosis Of Diabetic Retinopathy Using Machine Learning Techniques” ICTACT Journal On Soft Computing, July 2013, Volume: 03, Issue: 04
[18] Kranthi Kumar Palavalasa et, al “Automatic Diabetic Retinopathy Detection Using Digital Image Processing” International Conference on Communication and Signal Processing, 2018, India