International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Early prediction of diabetes using Feature Transformation and hybrid Random Forest Algorithm

B. Senthil Kumar, R. Gunavathi

Abstract: ***Diabetes is the most common chronic disease among the world. Early prediction of these will assist the physicians to provide the improved treatment. Machine learning approaches are widely used for predicting the disease at the earlier stage. However the selecting the significant features and the suitable classifier are still reduces the diagnosis accuracy. In this paper the PCA based feature transformation and the hybrid random forest classifier is utilized for diabetes prediction. PCA attempt to identify the best subset of transformed components that greatly improves the classification result. The system is compared with priori machine learning approaches to evaluate the efficiency of this work. The experimental result shows that the present study enhances the prediction accuracy.***

Keywords: ***Diabetes prediction, PCA, Random forest, machine learning, feature transformation***

I. INTRODUCTION

The chronic diseases are continual and their effects are permanent. The chronic condition affects the human's quality of life. Among the chronic diseases, diabetes is the most common disease present all over the world. The mortality rate is increase among the adults due to this chronic condition. People and government spend major budgets is spend on diabetes [1]. Diabetes was the fifth cause for mortality in women's and eighth cause for both sexes [2]. If it is continues large Number of people's quality of life will be affected. The only solution is to predict this choric condition in early stage to improve the treatment and reduce the morality rate due to diabetes. Diagnosing the diabetes in a still remains a challenging task among the researchers. The data mining technique play an important role in predict the diseases in early stage. Most of the studies [3-5] consider the machine learning algorithms for classifying the diseases. The issue in these techniques is to find the significant features from dataset and the suitable classifier. The proper selection of these two approaches is a challenging issue. This leads to the improper result with very low accuracy [6]. To overcome the above disadvantages the PCA based feature transformation is introduced in this study. The PCA tries to identify the best subset of transformed components which result in the improved accuracy. The ensemble pruning with back propagation neural network (BPNN) is introduced to improve the performance of Random forest. The hybrid

Revised Manuscript Received on July 22, 2019.
* Correspondence Author
B. Senthil Kumar*, Assistant Professor, Department of CA & IT, Sree Narayana Guru College, Coimbatore, (Tamil Nadu), India.
Dr. R. Gunavathi, Associate Professor and Head, Department of MCA, Sree Saraswathi Thyagaraja College, Pollachi (Tamil Nadu), India.

The remaining section of this study is organized as follows: section II describes the literature review of feature random classifier provides the efficient result with the transformed features.transformation and diabetes classification. Section III describes the dataset used in this approach and the proposed methodology. The experimental setup and the result are discussed in section IV. Eventually the summary of the study and the direction for future research is discussed in section V.

II. RELATED WORKS

The literature on feature selection approach for dimensionality reduction and classifying the disease using machine learning algorithms in medical dataset is discussed in this section. Mykola Pechenizkiy et al [7] performed the medical diagnostics process by applying the PCA based attribute transformation. The author proved that feature transformation provide better result than feature selection process for disease classification. The experiment was carried out with five different datasets like Diabetes, Heart, Cancer, Liver and Thyroid with 3-nearest neighbour classifier. TalhaMahboob Alam et al [2] introduced the Diabetes diagnosis model based on PCA based feature transformation technique. This study finds that diagnosis of diabetes has strong correlation with the two features body mass index and glucose level that was take out by aprior approach. The artificial neural network was used for prediction that achieved the 75.5% accuracy. Selvakuberan et al [8] presents the merits of data mining technique in diabetes prediction using ranker search method feature selection and Dagging classifier. The prediction model achieves 81.72 % accuracy. In addition the performance of this study is compared with 23 classification algorithm with weka tool. Deepti Sisodia and Dilip Singh Sisodia [9] applies the three best performing machine learning algorithms such as Decision tree, Support Vector Machine (SVM) and Naïve Bayes (NB) on pima dataset to predict the diabetes. Among them Naïve bayes obtain the great accuracy of 76.30%.

Berina Alić et al [10] utilizes the approaches Artificial Neural Networks and NB for both diabetes and cardiovascular diseases (CVD) classification. The NB achieves highest accuracy 99.51% and 97.92% on diabetes and CVD dataset respectively. HarleenKaur and VinitaKumari [11] applies the supervised machine learning technique on Pima dataset. The linear kernel SVM got 89 % accuracy compared to other approaches such as KNN, ANN, multifactor dimensionality reduction.

Tsehay Admassu Assegie, Pramod Sekharan Nair [12] investigates the performance of various machine learning methods on diagnosing the diabetes. Linear support vector machine got the highest accuracy of 78.39% accuracy compared to Gaussian naïve bayes and random forest.

From the literature it is clear that the prior research does not obtain the better accuracy on diabetes prediction. Therefore the proposed model presents a novel method with PCA based feature transformation and hybrid random forest classifier.

III. DATASET

Pima Indians diabetes dataset

Pima dataset is publically available in UCI machine learning repository which has 9 attributes with 768 patient records. The information is collected from the females with the age group of 21 years above in Pima Indian heritage. Among the 9 attributes the last one is the binary class attribute.

IV. PROPOSED METHOD

The early diabetes prediction model consists of two significant phase: feature transformation and Classification. The PCA method is applied in the present study for feature transformation and Hybrid random forest classifier is utilized for classification. The proposed architecture is shown in figure 1.

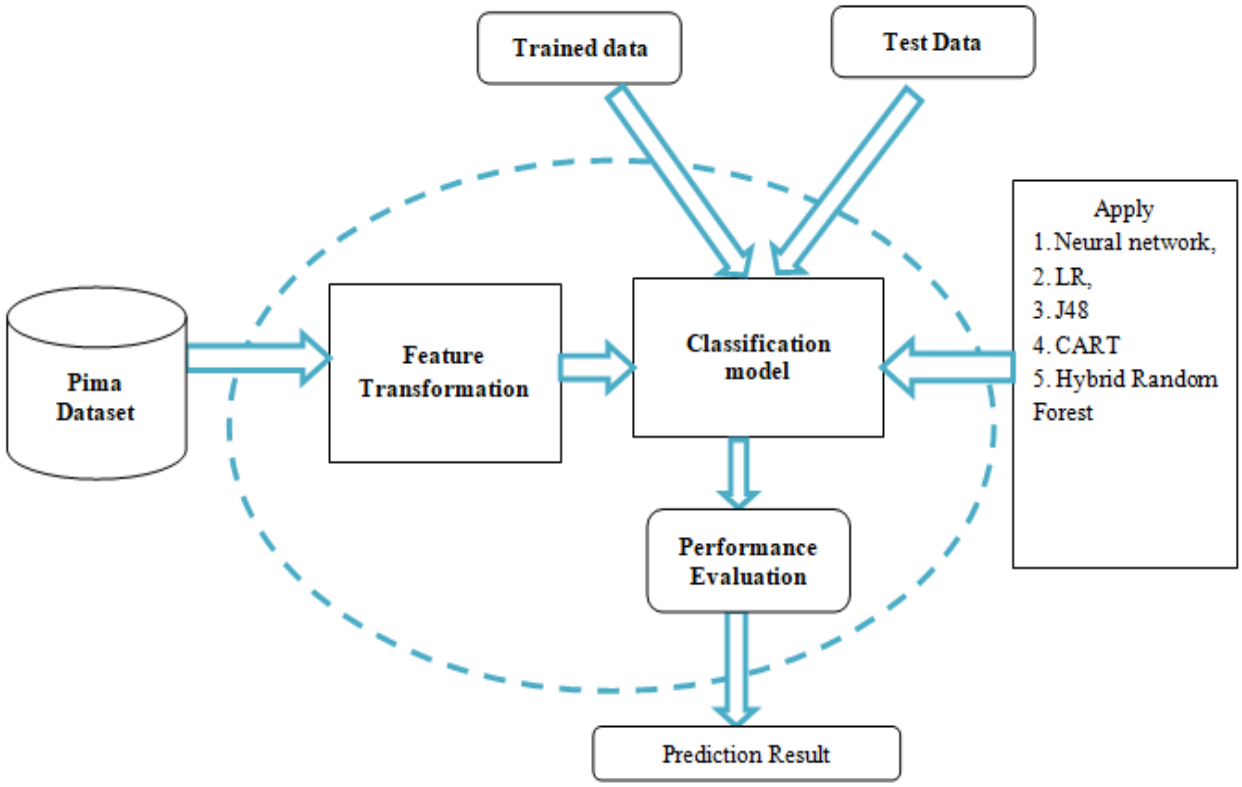

Figure 1. Architecture of the proposed model

Feature transformation using PCA

Feature transformation is performed to generate a set of features. The PCA is a most popular method which is adopted for feature transformation in the present study. This method is based on mining the axes on data that displays highest variability [7]. PCA provides the great support for supervised learning which spreads out the information's in news format.

A significant problem is to decide whether a PCA-based feature transformation approach is appropriate for a certain problem or not. Meanwhile the major objective of PCA is to mine new uncorrelated features, it is logical to present some correlation-based criterion with a possibility to define a threshold value. One of such criteria is the Kaiser-Meyer-Olkin (KMO) criterion that accounts for both total and partial correlation:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2} \quad (1)$$

Where R denotes the correlation matrix and A denotes the partial correlation matrix. $r_{ij} = r\left(x^{(i)}, x^{(j)}\right)$ represents the elements in R and a_{ij} represents the elements in A

$$a^2_{ij.X^{(i,j)}} = \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}} \quad (2)$$

Where $a^2_{ij.X^{(i,j)}}$ symbolizes the partial correlation coefficient for $x^{(i)}$ *and* $x^{(j)}$. Here i and j act as $X^{(i,j)}$(fixed controller) and R_{kl} (algebraic complement) in R.

When two attributes exchange a common factor with other attributes then their partial coefficient will be small that representing the distinctive variance they share. Likewise KMO value will be comes under the following factor

1. *if* a_{ij} *close to zero then* KMO value is close to one
2. *if* a_{ij} *close to one then* KMO value is close to zero

Popelínský [13] recommends using KMO with greater than 0.6 values for PCA. The study [13] showed that PCA with KMO> 0.5 will provide the great result.

Classification

During the training process only one type of rules can be learned by a single tree. In some cases the logic rule does not examine the connection among the input and output efficiently due to random subspace algorithm and bagging. These trees will reduce the overall performance of the system. Therefore ensemble pruning method has to be introduced to enhance the performance of the Random forest. In the proposed approach the back propagation (BP) neural network is utilized to develop a hybrid random forest. The neural weight in BP estimates the effects of inputs [14]. The Architecture of the proposed classification method is explained in figure 2.

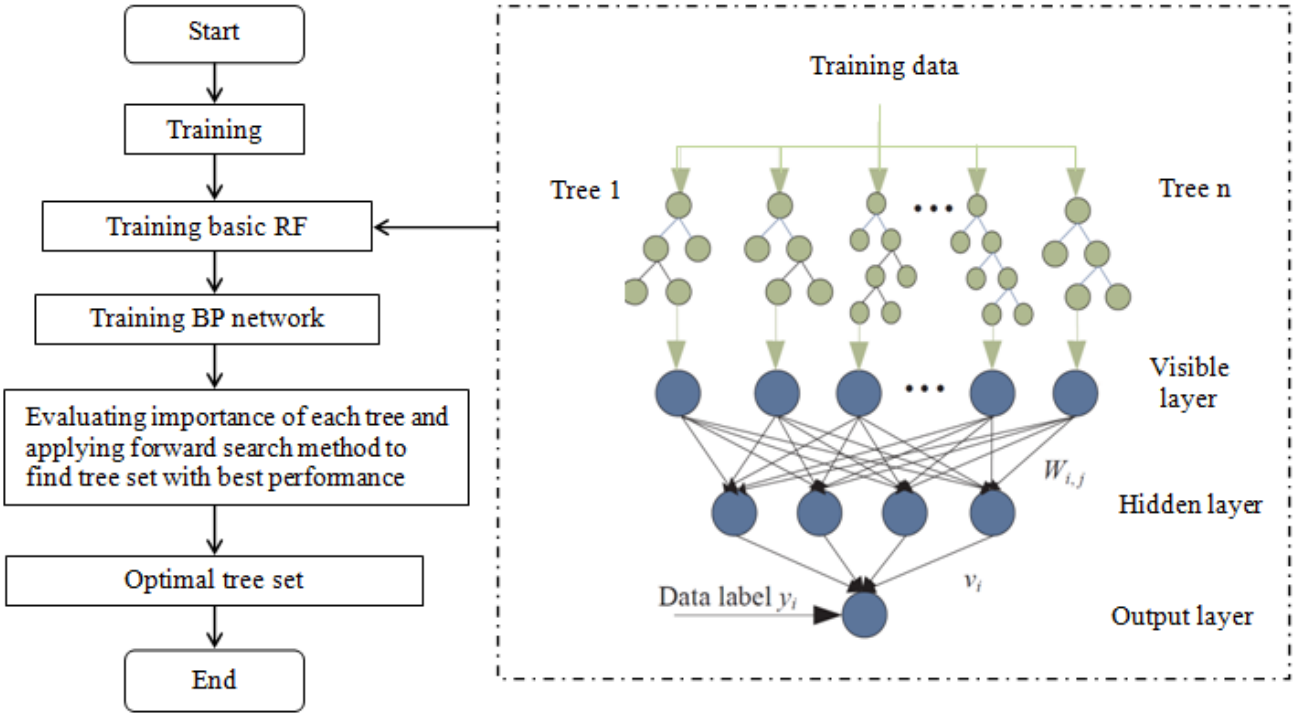

Figure 2. Architecture of the proposed hybrid Random forest

In this study, the output of each tree constitutes the input vector of the BP network, and the predicted data label is the output of the network. After training the network, the tree importance is evaluated. Based on the tree importance, a forward search method is applied to identify the tree set with optimum overall performance. The in-depth detail of the hybrid random forest approach is given in the previous work [15]. The algorithm of the classifier is given in table 1.

Hybrid RF Algorithm

Algorithm
Step 1. Upload the Training dataset
Step 2. Training the basic RF
Step 3. Training the BP network
Step 4. Calculating tree importance
Step 5. Outputting the optimal tree set

The Combination of PCA and Hybrid RF is worked well for the proposed research idea. The Feature transformation and the classification performed on those features provide the efficient result which is described in the following section.

V. EXPERIMENTAL RESULT

The experimental setup and performance result is discussed in this section. The experiment was carried out with Netbeans IDE and MySQl database. The publically available Pima dataset is utilized in the proposed prediction model. The feature transformation improves the accuracy compared to feature selection approach. The classification model provides better result on the transformed features.

The efficiency of the proposed model is evaluated by the performance metrics such as precision, recall, f-measures and accuracy. The table 1 presents the result of precision, recall and f-measures.

Table 1. Performance comparison of classification model

Algorithm	Precision	Recall	F-measures	Accuracy
Neural network	0.78	0.88	0.861	87.4
LR	0.817	0.924	0.894	90.7
J48	0.86	0.915	0.902	92.1
CART	0.74	0.876	0.824	89.1
Hybrid Random Forest	0.923	0.969	0.962	96.9

The performance of the proposed techniques on Pima dataset is shown in Figure 3 and 4. The combination of the proposed techniques (PCA and Hybrid RF) achieves the highest accuracy compared to other machine learning techniques.

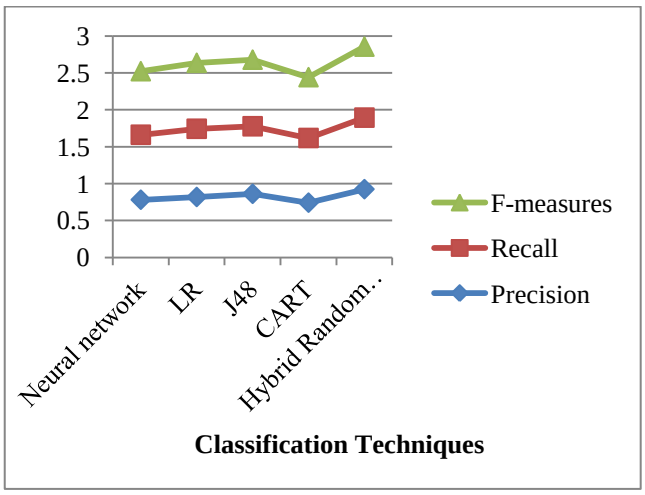

Figure 3. Performance Result Comparison

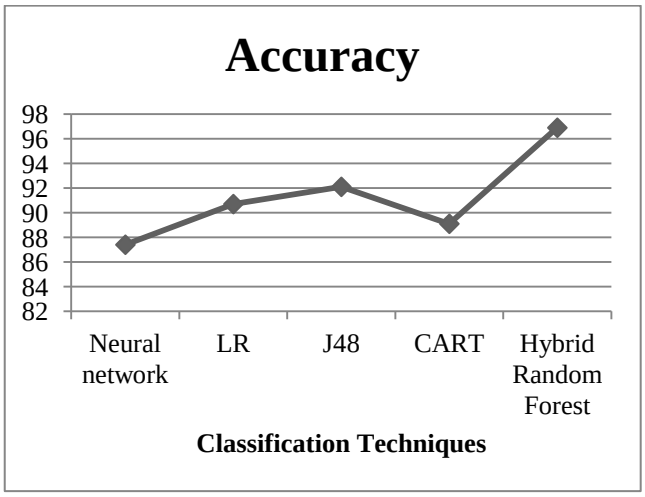

Figure 4. Accuracy comparison Chart

Figure 4 gives the accuracy comparison chart of proposed approach. The 96.9 is the greatest accuracy achieved by the present study. The other machine learning techniques are used in this study in order to evaluate the performance of the present work. Compared to those approaches the proposed study provides the better result for diabetes prediction.

VI. CONCLUSION

The present model proposed the early diabetes prediction model using feature transformation and classification techniques. Few prior classifiers for diabetes prediction are also discussed in terms of accuracy. The PCA method which is utilized for feature transformation provides the efficient result in reducing the components. The proposed hybrid random forest generates the great accuracy on the transformed feature. The present prediction model achieves 99% accuracy. In future the present study can be applied for other disease classifications like cancer, nervous related disease classification, liver disease classification etc.,

REFERENCES

1. Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 2017; 66:241–55.
2. Talha Mahboob Alam, Muhammad Atif Iqbal, YasirAli, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain Muhammad Awais Malik Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas, "A model for early prediction of diabetes", Informatics in Medicine Unlocked, Volume 16, 2019, 100204.
3. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science* 132 (2018): 1578-1585.
4. Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." *Applied Computing and Informatics* (2018).
5. Woldaregay, Ashenafi Zebene, Eirik Årsand, Ståle Walderhaug, David Albers, Lena Mamykina, Taxiarchis Botsis, and Gunnar Hartvigsen. "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes." *Artificial intelligence in medicine* (2019).
6. Shafenoor Amin, M., Kia Chiam, Y., Dewi Varathan, K., Identification of significant features and data mining techniques in predicting heart disease, *Telematics and Informatics* (2018).
7. Pechenizkiy, Mykola, Alexey Tsymbal, and Seppo Puuronen. "PCA-based feature transformation for classification: issues in medical diagnostics." In *Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems*, pp. 535-540. IEEE, 2004.
8. Selvakuberan, K., D. Kayathiri, B. Harini, and M. Indra Devi. "An efficient feature selection method for classification in health care systems using machine learning techniques." In *2011 3rd International Conference on Electronics Computer Technology*, vol. 4, pp. 223-226. IEEE, 2011.
9. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science* 132 (2018): 1578-1585.
10. Alić, Berina, Lejla Gurbeta, and Almir Badnjević. "Machine learning techniques for classification of diabetes and cardiovascular diseases." In *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1-4. IEEE, 2017.
11. Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." *Applied Computing and Informatics* (2018).
12. Tsehay Admassu Assegie, Pramod Sekharan Nair, "The Performance Of Different Machine Learning Models On Diabetes Prediction", International journal of scientific & technology research volume 9, issue 01, january 2020
13. B. Senthil Kumar *, Dr. R. Gunavathi, "An enhanced model for Diabetes prediction using Improves Firefly Feature selection and hybrid Random Forest Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

AUTHORS PROFILE

B. Senthil Kumar has completed his M.Phil. in Computer Science in Bharathiar University. He has 11 years of teaching experience and currently working as Assistant Professor, Department of CA & IT at Sree Narayana Guru College, Coimbatore. He has 8 years of research experience. He is now a Doctoral Student in Computer Science, Sree Saraswathi Thyagaraja College, Bharathiar University. His current field of research is Data Mining and Health Informatics. He guided 11 M.Phil scholars and published 22 papers in international journals.

Dr. R. Gunavathi has completed her Ph.D. in Computer science in Mother Teresa Women's University, Kodaikanal, and her research is on "Efficient Cluster head selection algorithms to improve the Quality of service in Mobile Ad hoc networks". She has 20 years of teaching experience and currently working as Associate Professor and Head, Department of MCA at Sree Saraswathi Thyagaraja College, Pollachi.She has 15 years of research experience. Her current research interest is in mobile ad hoc networks, Vehicular Ad hoc Networks and big data analytics. She has published around 30 research articles in the refereed International journals with good impact factor and also presented 25 research papers in the National and International level conferences. She has organized many National level Seminars, workshops and Conferences.