Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques

M G Dinesh¹ and D.Prabha²

¹ Research Scholar, Anna University, Chennai
² Asso Prof, Department of CSE, SKCET, Coimbatore,

E-mail: dineshbabu.mg@gmail.com

Abstract. Diabetes mellitus is a serious health issue in healthcare industry, which is a type of uncontrolled level of sugar. It is a chronic disease happened to the person who are having low insulin production and increase level of blood glucose because glucose is not properly utilized by body. In the medical field, predicting the correct diabetes is an important area that is under research to define a good predictive system to help the doctors to diagnose the disease. In the predictive system, feature selection plays on vital role to select the relevant feature for classification. There are several algorithms were applied on classification of diabetes data. In this proposed work, the features are transformed into high dimensional space before selection. So that the transformation of the features will give the better selection of attributes. With this effort, the proposed work implements the Kernel Principal Component Analysis for dimensionality reduction. KPCA will reduce the features space better than PCA. Once the features are transformed, the proposed work uses Genetic Algorithm to select the relevant and optimal features from the dataset. Then at the last Support Vector Machine is used as a classifier to classify the diabetes mellitus data. The proposed research on applying feature reduction before feature selection will reduce the irrelevant features that will improve the accuracy of the classification based on the selected relevant features. This proposed algorithm on diabetes mellitus data will compare with the existing algorithms to prove the effectiveness of the algorithm.

1. Introduction

Diabetes makes body energy less by preventing body from absorbing energy from in taking food. The food we eat is stored as glucose in liver by insulin hormone. This stored glucose is utilized by body when needed energy from body. The organ called pancreas behind our stomach is ductless gland used to produce insulin form its beta cells to convert food sugar into glucose. The two conditions of diabetes mellitus are:

1. Insulin does not secrete at all.
2. Secreted insulin is not utilized as – insulin resistance.

Every cells in human body needs energy to function. The sugar is main chemical component to give energy to cell. Sugar cannot be directly enter the cell, it needs key to get inside the cells. So insulin is key used to convert sugar into energy for cells. When insulin not works, sugar may be low or high in diabetes patients. There are two types of diabetes:

Type 1: beta cells are damaged and not produce the insulin in pancreas.
Type 2: insulin is generated, but cannot be utilized in body.
Sometimes gestational diabetes can be occurred for pregnant women. The main cause of the diabetes is stress, family history, pancreas damage, autoimmune deficiency, smoking etc.

WHO in 2016 estimated that 1.5 million deaths in 2012 are due to diabetes. It is also third high death causing agent in more countries. In India, day by day diabetes patients are increasing with high health risks. These diseases may cause complications like strokes, kidney failure, blindness, limbs amputations etc. worldwide diabetic patients in 2015 is 416 million whereas by 2040 it is estimated 650 million peoples will be affected by diabetes. India ranks second in diabetes index with 77 billion across worldwide 2019 after china.

Diagnosing the diseases accurately is main challenge among researchers. Delay in diagnosing causes more complication among the patients. More people doesn’t know if they affected by diabetes. It is necessary to introduce intelligent diagnosing methods to predict the diseases earlier and treat the patients.

1.1 Paper contribution

Diagnosing the diseases accurately is very difficult task. in our proposed model innovative technologies are merged to give accurate results. Feature selection method is important step in selecting the diseases based feature in test report accurately. The following are processed in our experiment.

1. The Kernel principal component analysis is chosen to eliminate irrelevant data and perform feature transformation.
2. Then the genetic algorithm is applied for effective feature selection to detect diseases based features.
3. SVM is best classifier appreciated by many researches for feature classification.

The rest of the paper is organized as giving review on existing methods in section 2. Proposed model is implemented and tested in section 3,4. Finally paper concludes with section 6.

2. Literature Review

The paper [1] discusses about the Principal Component Analysis (PCA), Kernel principal component analysis (KPCA) and independent component analysis (ICA) in terms of feature extraction using SVM. The original data is transformed to a new uncorrelated attributes in PCA. KPCA is a nonlinear version of PCA with the use of kernel function. This paper examine the sunspot data using SVM by the use of PCA, KPCA and ICA. Which lead to better result. Among the three methods, this work concludes that the KPCA performance is better compared to other two feature extraction methods.

In order to improve the accuracy of the existing diagnosis algorithms for the prediction of type 2 diabetes, the proposed algorithm [3] use the Goldberg’s genetic algorithm to select the features in the pre-processing step and use Multi Objective Evolutionary Fuzzy Classifier for classification on Pima Indians Diabetes Dataset. Feature selection by GA improve the accuracy of 83% since it reduce the features to 4 from 8.

This paper [6] uses decision tree for the prediction of Type II diabetes on Pima Indians Diabetes Data Set. The proposed work was divided as two steps such as pre-processing and prediction. Pre-processing stage consist of attribute identification and selection, missing values handling and discretization of the numerical data. Second stage of prediction use the decision tree model. This proposed work is implemented using Weka tool.

In this paper [7] Genetic Algorithm is used for feature selection and for classification to predict the selected features, J48graft Decision Tree algorithm has been used. This proposed work not only improves the accuracy but also reduce the cost, computation time and storage capacity. The results proves the proposed method with the accuracy of 74.78% and ROC 0.786 with the conclusion of GA and J48graft can be used for diagnosing the diabetes.

The paper [8] experimented Artificial Neural Network (ANN) with GA for feature selection on diabetic patients. GA is used to find the number of neurons in the hidden layer of ANN. This model is trained using Back Propagation Model and compared with Functional Link ANN, nearest Neighbor, K nearest neighbor, BSS- Nearest neighbor with backward sequential selection of feature, multiple feature subset to evaluate the accuracy of the suggested model. Paper [9] uses the Support Vector Machine (SVM) as
classifier to diagnose the diabetes disease on PIDD. They experimented and obtained the successful result on PIDD database. Paper [10] used Bayes network to predict the diabetes patients on PIDD data and proved that result was effective.

A hybrid approach based on UTA algorithm and Neural Network with two layer was proposed in paper [11] to predict the diabetes patient. This hybrid approach uses UTA algorithm for feature selection and neural genetic for prediction. Which is experimented on PIDD and predict of the diabetes patient with the accuracy of 87.46%. The proposed method also compared with other existing algorithms to evaluate the proposed algorithm. The Paper [12] uses ant colony based feature selection for classification. Ant Colony Optimization use the priori information of the attributes for feature selection. The performance of the classifier and the length of the feature selected are the important information in the proposed ACO work. So the features are selected based on the features having the shortest length and the classifier performance. The proposed work was experimented on ORL database of the Face Recognition System. The result proved that the proposed methodology is easy to implement and the performance is better than GA based feature selection methods.

The paper [13] proposed a kernel based RDA algorithm called kernel based one parameter regularized fisher discriminant (K1PRFD) based on one parameter RDA. The regularization and kernel are the two parameters used in this algorithm. Conjugate gradient method with scatter matrix is used to determine these parameters. This proposed work is implemented on face recognition to execute full feature space.

The paper [14] proposed highway traffic abnormal state detection using genetic and SVM. In this proposed work GA is used to select the SVM parameters and the two class SVM is changed to multi class SVM. Normally the state of the traffic is divided into three state such as no congestion, mild congestion and severe congestion. They proposed a multi classification algorithm called PCA-GA-SVM on Chongqing Wu Yu highway traffic flow data. PCA method is used to dimensionality reduction on weight basis then SVM kernel is used followed by GA for optimization. The proposed work is implemented using MATLAB and LIBSVM tool. This algorithm achieves high accuracy and detection rate compared to standard SVM.

The paper [15] proposed the feature selection based on principal component analysis and genetic algorithm. In this work, PCA has been used to transform the features followed by genetic algorithm to select features and for classification decision tree used. They use Canadian Institute of Cyber Security CICIDS2017 dataset for evaluation. They conducted experiment with PCA and without PCA before feature selection. And concluded the result as with PCA before feature selection will improve the classification accuracy compared to the normal one.

The aim of the proposed work [16] is to choose the best significant feature based prediction algorithm with the use of machine learning algorithm to find the best classification result. They proposed modified approach on selecting the relevant feature on diabetes dataset based on the correlation value. The experimented result shows that decision tree and random forest classification algorithm obtain specificity and sensitivity as higher. They suggested to transform the features from low dimension to high dimension will leads good classification result. This is the key point of our proposed work. In this paper [17], the authors proposed a system that use physiological model of blood glucose with support vector. The new system selects diabetes specialists at foreseeing blood glucose levels and can be used to predict right around a fourth of hypoglycaemic occurrences 30 min before of time. The comparing exactness is 42% and more false prediction are close to hypoglycaemic.

Generalized Discriminate Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) has been used to diagnose the diabetes disease has been proposed in paper [18]. And also authors proposed cascade learning system based on these techniques. They performed this in two steps. First step as preprocessing, GDA has been used to discriminant the features between the healthy and unhealthy patients. The proposed algorithm is compared with LS-SVM. The experimental results shows 82% of accuracy on classification than existing algorithm obtained only 78% of classification accuracy using cross validation on 10 fold. The robustness of this proposed algorithm has been experimented with accuracy, k fold cross validation and confusion matrix. With all metrics, the proposed algorithm obtains best result.
The paper [19] applied the feature selection algorithm on patient management system. Based on physiological and examination factors, the classification models were used to determine the diabetes disease of the patient. Supervised model construction (FSSMC) and optimization of Relief were used to select the important feature and after selecting the features three classification techniques such as Naïve Bayes, IB1 and C4.5 were applied to predict the patient status. FSSMC was proved a best feature selector for larger data sets.

The automatic prediction of type 2 diabetes by SVM has been proposed by [20]. They done the analysis on Pima Indian Diabetes data set as two parts. First part obtain the data recovery and feature selection. The selected features are the input to Multi-Layer Perceptron which is used as a classifier. This first part of the proposed work obtain 85% of accuracy. Second part they applied noise reduction method called k means for preprocessing followed by feature extraction. These extracted features are used to classify the data based on Random Forest and Logistic Regression classifiers. This second part obtain 77% accuracy. By comparatively with these two parts, data recovery with MLP obtain better level of accuracy than K means based noise reduction.

The feature selection algorithms such as K nearest Neighbor, K means and branch and bound etc., are discussed in paper [21]. They experimented each feature selection algorithm with the common diabetic data set and obtained the best algorithm for diabetes data feature selection. They concluded through the algorithmic survey that among the eight feature selection algorithms such as Naive Bayes, Support Vector Machine (SVM), C4.5, K-Nearest Neighbor, K-means, Branch and Bound, Randomized Hill Climb and Simulated Annealing, the branch and bound algorithm gives 100% accuracy and also it takes smaller amount of time for feature selection than other algorithms. This will leads to the prediction of diabetes.

Paper [22] use decision tree, random forest and neural network to predict diabetes mellitus on hospital physical examination data at china. 5 fold validation has been applied on the classification model. This model use 14 attributes and selected 68994 healthy pupils. For training phase, to avoid the data imbalance, the data are examined 5 times with the result as the average of these 5 experiments. The PCA and mRMR have been used for dimensionality reduction. While all attributes used, the random forest classification algorithm obtained best accuracy of 80%.

The study [23] with Bayesian and Decision Tress are used for Classification on diabetes dataset. With the help of Principal Component Analysis and Sequential forward selection algorithm, the 17 attributes are reduced to 4 attributes. They compared the classification result after applying the feature reduction algorithms. The dimensialityReduction on the data will improve the classification result and this will reduce the workload of machine learning. The experimented results prove that the decision tress classification after applying PCA will give better prediction of diabetes disease. Thus the proposed work gave importance to the dimension reduction part.

In terms of prediction and diagnosis of diabetes mellitus data and diabetic complications, the paper [24] presented the detailed study. Among the wide range of data mining classification algorithms 85% of the methods are supervised learning and 15% of unsupervised algorithms especially Association rule mining. This paper clearly explained the various machine learning algorithms are used to predict the disease on clinical data set which will leads to further research on Data mining on medical data.

The main contribution of the research [25] use leverage F Score feature selection algorithm and for classification they has been use fuzzy support vector machine. To train the dataset, SVM has been used that will generate the fuzzy rules and to classify the data, fuzzy inference process has been used. They experimented on Pima Indian Diabetes data set. The result obtained the accuracy of 89% to predict the diabetes mellitus patients. This research work suggested the as many fuzzy rules that will achieve sufficient accuracy will leads to further research on Diabetes Mellitus data prediction.
3. Proposed KPCA-GA-SVM Hybrid Model

3.1 Problem statement
Feature (Attribute) Selection is an optimization problem to select the subset of appropriate features for the purpose of classification. Selecting the relevant feature will lead to better classification result. So feature selection is an important step to be consider before classification. Even though feature selection and dimensionality reduction are grouped together, feature selection is selecting the relevant features without changing the attribute. Rather, dimensionality reduction is to transform the features into lower dimensional space.

In our case, we have used kernel principal component analysis for feature transformation followed by genetic algorithm to select optimal feature set and in the last, SVM as classifier for diabetes mellitus data prediction. The proposed approach shows that use of kernel principal component analysis before genetic algorithms improves the accuracy of the model with less number of features. The proposed model combines KPCA and GA with SVM. Comparative with PCA, KPCA and ICA, KPCA performs better reduction in terms of all dataset. With this point, our proposed work combines KPCA with GA for dimensionality reduction and feature selection.

3.2 Proposed Hybrid Architecture
In this model shown in Fig 1, we have been used KPCA along with GA for dimensionality reduction and feature selection and SVM for classification. The steps involve,
1. Feature reduction using KPCA
2. Feature Selection using GA
3. Classification Using SVM

![Figure 1. Proposed KPCA-GA with SVM architecture.](image-url)
1. KPCA Feature (Attribute) Reduction

KPCA is an extended version of PCA in nonlinear instance by the use of kernel method [1]. KPCA implicitly takes high order information of the original data with the use of kernel method. Which is generalize the PCA as nonlinear in KPCA. To achieve the generalization result as better, more number of principal component also extracted while using KPCA.

Algorithm 1: KPCA algorithm

Step 1: To transfer the original dataset \( T \) to the higher transformation, first step is to choose the kernel function \( k(x_i, x_j) \)

Step 2: Unlike PCA covariance matrix, KPCA will find the kernel matrix by applying the chosen kernel function to all pairs of attribute.

\[
K = T(x)T(x) \wedge T \text{ where,} \ (1)
\]

\[
\begin{bmatrix}
k(x_1, x_1)k(x_1, x_2)k(x_1, x_3) & \ldots & k(x_1, x_n) \\
k(x_2, x_1)k(x_2, x_2)k(x_2, x_3) & \ldots & k(x_2, x_n) \\
\vdots & \vdots & \vdots \\
k(x_n, x_1)k(x_n, x_2) & k(x_n, x_3) & \ldots & k(x_n, x_n)
\end{bmatrix}
\]

The value of each element in \( K = x_i, x_j \) (product of two vectors) in \( T \)

Step 3: calculate the mean of the transformed data and divide it by standard deviation

Step 4: find eigenvalues and eigenvectors of the matrix. And sort it based on eigenvectors in desc order.

Step 5: concatenate the first \( m \) Eigen vector with the number of reduction \( m \) as one matrix. And calculate the product of the matrix with the original data. We will get the reduced dataset for further processing.

2. Genetic Algorithm – Feature Selection

With the types of feature selection algorithms such as filter, wrapper and embedded, the bigger dataset use wrapper methods to improve the accuracy. Once the data transformed to high dimension, the next step feature selection is performed on the transformed data. In this work, genetic algorithm based feature selection is performed on the high dimensional transformed data. Which is the heuristic based method under wrapper feature selection method. Genetic algorithm used in various application mainly suitable for medical data [3]. Hence, our proposed work feature selection on the diabetes mellitus data set, genetic algorithm is used after KPCA reduction. Genetic algorithms have the components such as fitness function, chromosomes population, Selection, crossover and mutation. The algorithm steps are,

Algorithm 2:

Step 1: Initialize the individual in the population

Step 2: Calculate the Fitness function of the individuals in the population

Step 3: till find the fittest result in the population repeat selection, crossover and mutation.
3. Support Vector Machine – Classification
Since, SVM has been used for classification in the area of machine learning and feature selection, our proposed work implement the SVM classifier for classification. SVM is a supervised classification technique [4]. The SVM divide the data set into two classes through the hyper plane. First identifies the optimal hyper plane in the data and then mapping the attributes into the boundaries specified by the model. The training space of SVM assigns the new samples into the any one of the divided classes. The Fig 2 shows the SVM classification on diabetes mellitus data

3.3 Proposed algorithm (KPCA-GA-SVM)
As we stated earlier, the diabetes mellitus prediction of our proposed work focuses on KPCA for feature reduction, GA for Feature Selection and SVM for classification. The proposed algorithm flow is sated in Fig 3.
Algorithm 3:

Step 1: Input the diabetes mellitus data set
Step 2: Apply KPCA for feature reduction. So that based on the kernel matrix and product operation of all the attributes. Based on eqn 1 and KPCA algorithm 1, the diabetes data attributes are reduced to high dimensional space. The original and transformed dataset are shown in Fig 4.

---

**Figure 3.** Proposed KPCA-GA-SVM algorithm work flow.

1. Genetic manipulation
2. Regenerate populations

---

Diabetes
Non Diabetes
Step 3: generate the chromosomes population randomly. Evaluate the Fitness Function on attributes of the population using eqn 2

\[
\text{Fitness } f(x) = \frac{\text{fitness of an individual attribute } (i)}{\text{sum of fitness of all attributes } (x_i, x_j)}
\]  

(2)

Step 4: While the termination condition not satisfied, regenerate the population based on the fitness value and repeat selection, crossover and mutation repeatedly on the new data attributes.

Step 5: while termination condition met, then best attribute is found for classification. Perform classification based on SVM on the selected attribute.

Step 6: the classification results predict the correct diabetes patient. Thus using the proposed KPCA-GA-SVM improve the accuracy of the prediction.

4. Results and Discussions

This section describes about the result obtained by using our proposed KPCA-GA-SVM for diabetes prediction. To prove the proposed algorithm accuracy, this algorithm is compared with the existing approaches such as without dimensionality reduction, and neural network algorithms and some standard feature selection algorithms.

4.1 Diabetes Mellitus Data set Description

For evaluation, the diabetes data set called Pima Indians Diabetes Data Set is collected from UCI repository [5]. The main aim of our work is to predict the patient have diabetes or not. The data set consist of 768 samples with 15 attributes and 1 class attribute which consist of binary value 0 or 1. The class attribute 0 means the patient diabetes result is negative and 1 means the patient diabetes result is positive. The data set features (attributes) are shown in Table 1.
Table 1. Pima Indian diabetes dataset attributes.

| Attribute No. | Attribute                                                   |
|--------------|-------------------------------------------------------------|
| 1            | Number of times pregnant (NTP)                             |
| 2            | Plasma Glucose concentration (PGC)                         |
| 3            | Diastolic Blood Pressure (mmHg) (DBP)                      |
| 4            | Triceps Skin- Fold thickness (mm) (TSFT)                   |
| 5            | 2-h Serum insulin (mu U/mL) (H2SSI)                        |
| 6            | Body Mass Index (kg/m2) (BMI)                              |
| 7            | Diabetes Pedigree Function (DPF)                           |
| 8            | Age                                                        |
| 9            | Class 0 and 1 (Diagnosis of type 2 diabetes disease)       |

4.2 Experimental Results
The proposed algorithm is experimented in python Scikit learn. The algorithmic formation is declared as: number of principal components for KPCA is 80, number of chromosome generation for GA as 12 with a population size of 150 for each generation, mutation probability as 0.3 and crossover probability as 0.5.

4.3 Evaluation criteria
The classification result is formulated using accuracy based on True Positive, False Negative, True Negative and False positive.

True Negative (TN): The number of samples that are not belonged to the desired class and the network is detected correctly.

True Positive (TP): The number of samples that are belonged to the desired class and the network is detected correctly.

False Negative (FN): The number of samples that are not belonged to the desired class and the network is detected correctly.

False Positive (FP): The number of samples that are belonged to the desired class and the network is not detected correctly.

\[
\text{Accuracy} \ (\%) = \frac{TN + TP}{TN + FN + TP + FP} \quad (3)
\]

We experimented two level of evaluation with the diabetes dataset as KPCA-GA-SVM without dimensionality reduction as GA-SVM approaches. The result for our proposed algorithm is stated in Table 2. The result shows that, after the attributes are transformed using KPCA, the selection of features through GA will reduced (4,5,6,7) which leads to good classification result Rather without dimensionality reduction, the features selection count is little increased which leads to the classification accuracy lesser than the proposed one.

Table 2. Pima Indians Diabetes Data Set experimental results.

| Approach            | Total Number of Feature | Accuracy with All Feature | Number of Reduced Feature | Accuracy with Reduced Feature |
|---------------------|-------------------------|---------------------------|---------------------------|------------------------------|
| KPCA-GA-SVM         | 8                       | 98.23                     | 4                         | 99.53                        |
| GA-SVM              | 8                       | 97.45                     | 5                         | 98.89                        |

From The table 2, we proved that our proposed KPCA-GA-SVM obtain accuracy of 99.53% and also reduced feature size compared to GA-SVM of 98.79% accuracy. The result for this table 2 is depicted in Fig 5.
Next metrics used to evaluate the proposed algorithms are Precision, Recall, Fall out, F measure and ROC.

**Precision:** it is the number of samples that are correctly classified as diabetic from the total number of samples in the data fetched.

\[
Precision = \frac{TP}{TP+FP} \tag{4}
\]

**Recall:**
It is the measure that correctly classified the samples that are actually have the diabetic disease to the total number of instances in the data set.

\[
Recall = \frac{TP}{TP+FN} \tag{5}
\]

**F - Measure:** It is the average measure of Precision and Recall which is calculated using (6)

\[
F - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}
\]

**ROC:** ROC (Receiver operating curve graph) is an effective method of evaluating the performance of the classifier for the diagnostic tests.

\[
ROC = \frac{1}{2} \left( \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \tag{7}
\]

The evaluated results are shown table 3. From the table, the proposed methodology results shows bettercompare to the feature selection without KPCA. Which is depicted in Fig 6.

**Table 3.** Performance Metrics on PIDD data using KPCA-GA-SVM.

| Algorithms      | Precision | Recall | F-Measure | ROC  |
|-----------------|-----------|--------|-----------|------|
| KPCA-GA-SVM     | 92.45     | 91.4   | 91.925    | 0.98 |
| PCA-SVM         | 89.34     | 87.5   | 88.42     | 0.78 |
Figure 6. Performance Metrics on PIDD data using KPCA-GA-SVM.

The ROC (Receiver Operator Characteristic) curve used to represent the performance of binary classification techniques. Since our dataset is binary classification, our proposed approach diagnostic test is done with ROC curve (0.98) in terms of false positive and True positive rate are shown in Fig 7.

Figure 7. ROC curve of KPCA-GA-SVM.

To evaluate the accuracy of our proposed algorithm with other benchmarks data mining and machine algorithms, KPCA-GA-SVM on diabetes dataset is compared with the hybrid model of fuzzy system and regression tree [7], Decision tree [6], Artificial neural network [8], SVM [9], Bayesian network [10]. The performance compassion is shown in Fig 8. The analysis result clearly shows that our proposed KPCA-GA-SVM on diabetes data set give better accuracy of 98.56% compared to other approaches with the respective percentage of 87.89%, 85%, 79%, 88.76%, 76.34% and 95%.
Further to evaluate the classification effectiveness three more metrics are used. They are sensitivity, specificity and Matthews’s correlation coefficient (MCC) based the criteria of the accuracy metrics. The experimented results are shown in Table 4. Which is pictorially represented in Fig 8.

**Figure 8.** Comparison Evaluation of Various Algorithms with Proposed on Diabetes data set.

The Fig 9, depicts that our proposed algorithm performs better in terms of sensitivity (96.4%), specificity(94%), accuracy(97.3%) and MCC (89.3%) compared to other classification algorithms.

**Table 4.** Comparison Evaluation through various metrics.

| APPROACHES                             | ACC     | SN      | SP      | MCC     |
|----------------------------------------|---------|---------|---------|---------|
| The hybrid model of fuzzy system and regression tree | 82.77   | 0.7923  | 0.7773  | 0.6542  |
| Decision tree                          | 0.7634  | 0.7678  | 0.7454  | 0.576   |
| Artificial neural network              | 0.7677  | 0.7865  | 0.7678  | 0.5459  |
| SVM                                    | 0.7377  | 0.7465  | 0.7578  | 0.5459  |
| Bayesian network                       | 0.7295  | 0.7087  | 0.7463  | 0.4697  |
| GA-SVM                                 | 0.8778  | 0.8037  | 0.8237  | 0.7538  |
| KPCA-GA-SVM                            | 0.973   | 0.9689  | 0.9403  | 0.8934  |
The experimented result with the proposed approach on diabetes mellitus data shows the better accuracy on classification because the features are transformed using KPCA before selection. Due to this, the feature selection Genetic Algorithm selects only the relevant attributes for classification. This algorithm is compared with existing algorithms, and the results also shows our proposed work achieves good accuracy on prediction which predict the correct diabetes patients.

5. Conclusion and Future Work

This is a new hybrid method developed from existing Efficiency proved algorithms. Kernel based PCA is used in reduction of duplicate data in the data set. This helps to improve the efficiency in detecting the diseases from accurate results. The proposed work uses genetic algorithm to detect the accurate feature from the data set. Then the SVM classifier finally used to classify the actual feature infected by diseases. The accuracy of diagnosing the diabetes is high when compared with existing models. In future various image processing algorithms can be used in diagnosing diabetes in early stages. We can also focus on food based diabetes prediction in our future research work.

References

[1] L.J. Caoa;∗, K.S. Chuab, W.K. Chongc, H.P. Leea, Q.M. Gud, “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine”, Elsevier, Neurocomputing 55 (2003) 321 – 336.

[2] https://ig.opengenus.org/kernal-principal-component-analysis/

[3] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-5, doi: 10.1109/ICCNI.2017.8123815.

[4] Chapelle O, Hafner P, Vapnik V. Support vector machines for histogram-based image classification. IEEE Trans Neural Netw. 1999;10(5):1055–64.
[5] V. Sigillito, Data set Pima Indians Diabetes collected (1990), https://archive.ics.uci.edu/ml/datasets/Pima+Indians +Diabetes, Accessed dec 8, 2015.
[6] A. AlJarullah, “Decision Tree Discovery for the Diagnosis of Type II Diabetes”, International Conference on Innovations in Information Technology, 2011
[7] M. Seera and Ch. P. Lim, “A hybrid intelligent system for medical data classification”, Expert Systems with Applications, vol.41, pp. 2239–2249, 2014.
[8] M. Pradhan and R. K. Sahu, “Predict the onset of diabetes disease using Artificial Neural Network (ANN)”, International Journal of Computer Science & Emerging Technologies, vol. 2, pp 2044-6004, 2011.
[9] V. A. Kumari and R. Chitra, “Classification Of Diabetes Disease Using Support Vector Machine”, International Journal of Engineering Research and Applications, vol. 3, pp. 2248-9622, 2013.
[10] Y. Guo, G. Bai and Y. Hu, “Using Bayes Network for Prediction of Type-2 Diabetes”, The seventh International Conference for Internet Technology and Secured Transactions, 2012.
[11] Dadgar, S. M. and MostafaKaardaan. “A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes.” (2017).
[12] H. R. Kanan and K. Faez, “An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system,” Appl. Math. Comput., vol. 205, no. 2, pp. 716–725, 2008.
[13] W.S. Chen, P.C. Yuen, J. Huang, D.Q. Dai, Kernel machine-based one-parameter regularized fisher discriminant method for face recognition, IEEE Transactions on Systems, Man, and Cybernetics-Part B 35 (4) (2005) 659–669.
[14] Z. Min, L. Yanlei, S. Dihua and C. Senlin, "Highway traffic abnormal state detection based on PCA-GA-SVM algorithm," 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, 2017, pp. 2824-2829, doi: 10.1109/CCDC.2017.7978993.
[15] Rahul Adhao, VinodPachghare, “Feature selection using principal component analysis and genetic algorithm”, February 2020Journal of Discrete Mathematical Sciences and Cryptography 23(2):595-602DOI: 10.1080/09720529.2020.1729507
[16] Sneha, N., Gangil, T. “Analysis of diabetes mellitus for early prediction using optimal features selection”. J Big Data 6, 13 (2019). https://doi.org/10.1186/s40537-019-0175-6
[17] Kevin P, Razvan B, Cindy M, Jay S, Frank S. “A machine learning approach to predicting blood glucose levels for diabetes management. In: Modern artificial intelligence for health analytics” Papers from the AAAI-14, 2014.
[18] Polat K, Güneş S, Arslan A. “A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine”. Expert Syst Appl. 2008;34(1):482–7.
[19] Huang Y, McCullagh P, Black N, Harper R. “Feature selection and classification model construction on type 2 diabetic patients' data.” ArtifIntell Med. 2007 Nov;41(3):251-62. doi: 10.1016/j.artmed.2007.07.002. Epub 2007 Aug 17. PMID: 17707617.
[20] S. Y. Rubaiat, M. M. Rahman and M. K. Hasan, "Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection," 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2018, pp. 1-6, doi: 10.1109/CIET.2018.8660831.
[21] R. Lomte, S. Dagale, S. Bhosale and S. Ghodake, "Survey of Different Feature Selection Algorithms for Diabetes Mellitus Prediction," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697452.
[22] Quan Zou, Kaiyang Qu, YameiLuo, Dehui Yin, Ying Ju and Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques”, Machine Learning Advanced Dynamic Omics Data Analysis for Precision Medicine, Front. Genet., 06 November 2018 | https://doi.org/10.3389/fgene.2018.00515.
[23] Hasan Akan, Ugur Demirok, Niyazi Kilic, “The effect of feature reduction techniques on diagnosis of diabetes”, April 2012, DOI: 10.1109/SIU.2012.6204687.

[24] Ioannis Kavakiotis, Olga Tsav, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal Volume 15, 2017, Pages 104-116, https://doi.org/10.1016/j.csbj.2016.12.005.

[25] Rian Budi Lukmantoa, Suharjitoa,*, Ariadi Nugrohoa, Habibullah Akbar, “Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine”, Elsevier, 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019, Procedia Computer Science 157 (2019) 46–54.