IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 8, 2019, accepted July 23, 2019, date of publication July 30, 2019, date of current version August 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931956

A Linear Model Based on Principal Component Analysis for Disease Prediction

H. ROOPA[1] **AND T. ASHA**[2]
[1]Department of Information Science and Engineering, Bangalore Institute of Technology, Bengaluru 560004, India
[2]Department of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru 560004, India

Corresponding author: H. Roopa (roopatejas@gmail.com)

ABSTRACT Various classification methods are applied to predict different diseases, such as diabetes, tuberculosis, and so on, in medical field. Diagnosis of diabetes can be analyzed by checking the level of blood sugar of patient with the normal known levels, blood pressure, BMI, skin thickness, and so on. Several classification methods have been implemented on diabetes. In this paper, the main aim is to build a statistical model for diabetes data to get better classification accuracy. Extracted features of diabetes data are projected to a new space using principal component analysis, then, it is modeled by applying linear regression method on these newly formed attributes. The accuracy obtained by this method is 82.1% for predicting diabetes which has reformed over other existing classification methods.

INDEX TERMS Principal component analysis, linear regression model, diabetes.

I. INTRODUCTION

Analysis of diseases is a tough task in medical field. Diagnosing diabetes [2] can be understood by checking the blood sugar level with normal desired level. In this manuscript we present a statistical diabetes disease prediction model where Principal Component Analysis (PCA) is applied to extract attributes of Pima Indian Diabetes Data (PIDD) to a new feature space. These attributes are then modeled using linear regression model [8] to predict diabetes.

Attributes of PIDD are inspected at different angles to obtain required information for processing data. So, feature extraction is a major step in examining PIDD. The work concentrates on retrieving feature values from PIDD to a new feature space by employing PCA method. These new set of feature values are inspected for their importance and relevance, and are subjected for data mining methods like LRM to classify the given data for predicting diabetes disease.

PCA is reduction method which considers the PIDD as set of rows representing characteristics in a high dimensional space and all rows are put up to a directions which represents the best set of features. Here for original set of attributes of PIDD, this transformation is applied to obtain an axis that contains principle eigenvector where all the points of all observations of each feature are spread out. Maximized variance of data can be found on this axis. When considering second eigenvector, observe that axis along which the variance of distance from first axis is greatest and so on. Small number of eigenvectors is represented by matrix of points, then to minimize the root mean square error, approximation of data is performed for the given number of columns in the matrix consider. Thus the original features of PIDD are approximated with fewer dimensions which are an overall of original PIDD.

Model building provides a good fit to any set of data. Linear statistical model estimate the unknown dependent PIDD feature value from the known independent PIDD feature values. The representation of relationship between dependent PIDD image feature and set of independent multiple PIDD features are known as regression analysis.

The work is explained as follows, section 2 describes about previous work on PIDD, proposed methodology is explained in section 3, section 4 discusses about findings and section 5 gives about work's conclusion.

II. RELATED WORK

Polat *et al.* [1] proposed a Least Square Support Vector Machine (LS-SVM) classification method to obtain an accuracy of 79.16% which improvised over previous classification methods. Generalized Discriminant Analysis (GDA) was used at preprocessing stage for discriminating variables of PIDD and then LS-SVM technique was applied on these variables for classifying the disease.

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Zhang.

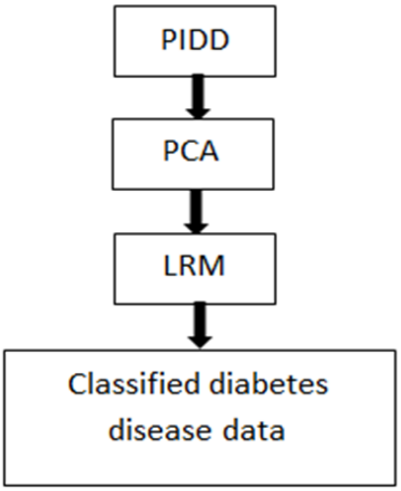

FIGURE 1. Proposed steps involved in feature extraction and modeling of PIDD.

Seera and Lim [3] proposed Fuzzy Min–Max neural network, Regression Tree and Random Forest (FMM-CART-RF) combined hybrid classification method that achieved an accuracy of 78.39% for PIDD.

Sa'di *et al.* [4] classified PIDD using various data mining algorithms and analyzed that Naïve Bayes performed well than RBF network and J48 with an accuracy of 76.95%.

Bansal *et al.* [5] proposed an evolutionary method where feature selection of PIDD is obtained by Particle Swarm Optimization (PSO) method and then k-Nearest Neighbor (KNN) classification technique is applied on these features to achieve an accuracy of 77%.

Choubey *et al.* [6] applied Genetic Algorithm (GA) for variable selection on PIDD. Then implemented Naive Bayes (NB) on these selected variables for classifying the diabetes disease to obtain an accuracy of 78.69%.

III. PROPOSED METHODOLOGY

The proposed method includes extraction of new group of features from PIDD by employing PCA so that the values are inspected for their importance and relevance, and are subjected for data mining methods like Linear Regression Model (LRM) to classify the given data for predicting diabetes disease. The model is illustrated as shown in Figure 1.

A. INPUT DATA

The PIDD has been taken from UCI of machine learning repository [9]. In this dataset female patient from Phoenix and Arizona aged about 21 years are considered. The PIDD consists of '8' actual variables and '1'class variable. There are totally 768 instances out of which 268 instances have class variable value '1' and 500 instances have class variable value '0' respectively. The patients are classified as diabetes or normal using binary valued variables. If the binary response variable represents '1' means ''positive for diabetes'' and '0' means ''negative for diabetes''. The attributes of PIDD are presented in Table 1.

TABLE 1. The eight attributes of PIDD with the class variable.

No.	Variable	New Variable	Value
1.	Pregnancies	Comp1	Integer
2.	Glucose	Comp2	Integer
3.	Blood Pressure	Comp3	Integer
4.	Skin Thickness	Comp4	Integer
5.	Insulin	Comp5	Integer
6.	BMI	Comp6	Numeric
7.	Diabetes Function	Comp7	Numeric
8.	Age	Comp8	Integer
9.	Class	Class	Integer (0 or 1)

B. FEATURE EXTRACTION [7] OF PIDD BY APPLYING PCA

For the given numeric data matrix containing the feature values of PIDD apply principal component analysis to project the features to a new space.

The steps involved in feature extraction are

1) From each dimensions of PIDD subtract the mean, which produces a data set whose mean is zero.
2) Calculate the variance matrix between two separate dimensions of PIDD.
3) Find the Eigen vector and Eigen values of the matrix obtained in step2.
4) Construct the feature vector and take transpose of it. Then multiply it with original PIDD to obtain new set of features projected to new space.

C. LINEAR REGRESSION MODEL (LRM)

Regression analysis predict dependent variable value 'y' based on the set of independent variable values $x_1, x_2, x_3, \ldots, x_k$.

The multiple independent variables analysis, linear regression model [10] is represented by equation (1)

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon \quad (1)$$

where 'y' is the class variable of PIDD data,

b_0 is a 'y' intercept,

ε is an error term and

$b_1, b_2, b_3, \ldots\ldots\ldots, b_k$ are coefficients of $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_k$, respectively.

The dependent variable 'y' of PIDD is the class variable, which is predicted using equation (1). 'y' is based on the

set of independent variable values $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_k$ represented by values of comp 1, comp 2, comp 3, comp 4, comp 5, comp 6, comp 7 and comp 8 respectively.

The fitted line of equation (1) is calculated by principle of least square method where b_0, b_1, b_2, $b_{3,\ldots\ldots\ldots}$, b_k are chosen in such a way that the Sum of Squares of Error (SSE) is minimum.

Consider equation (1), For n = 768 observation,

Let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{738} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} x_{21} \ldots & x_{81} \\ \vdots & \ddots & \vdots \\ 1 & x_{1786} \ldots & x_{8768} \end{bmatrix}$$

$$\hat{b} = \begin{bmatrix} \hat{b_0} \\ \hat{b_1} \\ \hat{b_2} \\ \vdots \\ \vdots \\ \hat{b_k} \end{bmatrix}$$

$\hat{b} \rightarrow$ least square estimates of b_0, b_1, …, b_k of linear model.

Least square matrix is obtained by

$(x^1x)\hat{b} = x^1y$

where $x^1 \rightarrow$ is transpose of x matrix.

$(x^1x) \rightarrow$ coefficient matrix of least square estimates of $\hat{b_0}, \hat{b_1} \ldots\ldots\ldots \hat{b_k}$.

$x^1y \rightarrow$ gives matrix of constants.

Therefore least square solution is obtained by $\hat{b} = (x^1x)^{-1}.x^1y$

Substitute $\hat{b}$ in equation (1) to get the final fitted line.

D. OUTPUT DATA

The model classifies features of PIDD as diabetic or normal.

IV. RESULTS

The attributes of PIDD is projected to a new space using PCA. Then LRM is applied on the components of PIDD. The results of the model are presented in Figure 2.

The residuals error between the prediction of LRM and actual results is smaller ranging between −1.0059 and 1.2930. And the median value is closer to zero. F-statistics shows that model has at least one variable that is significantly different than zero. The variables comp1, comp2, comp5, comp6, comp7 and comp8 are more significant indicated by *** and ** when compared to comp3 and comp4 respectively, which are less significant.

The components of PIDD are partitioned, where 80% are considered for training the model and 20% of data for testing the model respectively. These data are classified using

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0059 -0.2937 -0.1120  0.3357  1.2930

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3470905  0.0146867  23.633  < 2e-16 ***
Comp.1      -0.0006238  0.0001266  -4.926 1.03e-06 ***
Comp.2      -0.0069757  0.0004813 -14.492  < 2e-16 ***
Comp.3      -0.0003614  0.0007452  -0.485 0.627897
Comp.4      -0.0025501  0.0010449  -2.440 0.014895 *
Comp.5       0.0063874  0.0013864   4.607 4.79e-06 ***
Comp.6      -0.0072193  0.0021763  -3.317 0.000953 ***
Comp.7       0.0182008  0.0052844   3.444 0.000604 ***
Comp.8      -0.1469275  0.0459216  -3.200 0.001434 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4067 on 759 degrees of freedom
Multiple R-squared:  0.2806,    Adjusted R-squared:  0.273
F-statistic: 37.01 on 8 and 759 DF,  p-value: < 2.2e-16
```

FIGURE 2. LRM results of PIDD.

```
Confusion Matrix and Statistics

          actual
predicted   0   1
        0 351  99
        1  46 110

               Accuracy : 0.7607
                 95% CI : (0.7247, 0.7942)
    No Information Rate : 0.6551
    P-Value [Acc > NIR] : 1.160e-08

                  Kappa : 0.4367
 Mcnemar's Test P-Value : 1.572e-05

            Sensitivity : 0.8841
            Specificity : 0.5263
         Pos Pred Value : 0.7800
         Neg Pred Value : 0.7051
             Prevalence : 0.6551
         Detection Rate : 0.5792
   Detection Prevalence : 0.7426
      Balanced Accuracy : 0.7052

       'Positive' Class : 0
```

FIGURE 3. Performance of training data.

LRM. The actual and predicted values of PIDD modeled by LRM are given by confusion matrix. The confusion matrix and statistics of training and testing data of PIDD are given in Figure 3 and Figure 4 respectively.

The confusion matrix of training data as given in Figure 3 depicts that 614 instances out of 768 instances of PIDD are considered for training. 351 observations are correctly predicted negative for diabetes and 110 observations are correctly predicted positive for diabetes. So, 461 observations have been correctly classified by LRM. 46 observations represent incorrect prediction of LRM that resulted positive for normal person and 99 observations represent incorrect prediction of LRM that resulted negative for diabetic person. So, 145 observations have been wrongly classified by LRM.

The confusion matrix of testing data as given in Figure 4 depicts that 154 out of 768 instances of PIDD are considered for testing. 96 observations are correctly predicted negative for diabetes and 37 observations are correctly predicted positive for diabetes. So, 133 observations have been correctly classified by LRM. 7 observations represent incorrect prediction of LRM that resulted positive for normal person and 22 observations represent incorrect prediction of LRM that resulted negative for diabetic person. So, 29 observations have been wrongly classified by LRM.

```
Confusion Matrix and Statistics

          actual
predicted  0  1
        0 96 22
        1  7 37

               Accuracy : 0.821
                 95% CI : (0.7531, 0.8767)
    No Information Rate : 0.6358
    P-Value [Acc > NIR] : 1.956e-07

                  Kappa : 0.5913
 Mcnemar's Test P-Value : 0.00933

            Sensitivity : 0.9320
            Specificity : 0.6271
         Pos Pred Value : 0.8136
         Neg Pred Value : 0.8409
             Prevalence : 0.6358
         Detection Rate : 0.5926
   Detection Prevalence : 0.7284
      Balanced Accuracy : 0.7796

       'Positive' Class : 0
```

FIGURE 4. Performance of testing data.

TABLE 2. The performance evaluation of training and testing data using PCA-LRM on PIDD.

Performance metrics	Performance of training data	Performance of testing data
Precision	0.78	0.813
Recall	0.884	0.932
F-Measure	0.828	0.868
Accuracy	76.07	82.1
Kappa statistics	0.4367	0.5913
P-Value	1.160e-08	19.956e-07
Sensitivity	0.8841	0.9320
Specificity	0.5263	0.6271

The performance of LRM can be analyzed by its accuracy in prediction of disease. The implementation of LRM on PIDD uses the concept of probability to reveal prediction of diabetic or nondiabetic patient. The default probability value considered is 0.5 which is the cutoff value. If the probability value falls below 0.5 then the possible LRM prediction is negative for diabetes otherwise if it is above 0.5, LRM predicts the patient to be diabetic.

The overall performance of the model on PIDD is represented in Table 2.

From PIDD, 80% of data is considered for training the data. The snapshot of performance curve of training data of PIDD is shown in Figure 5. The performance curve of LRM is analyzed with respect to accuracy and cutoff value used in disease prediction. Figure 5 depicts that at cutoff value 0.5, LRM prediction accuracy raises above 70%. LRM prediction accuracy falls low when the cutoff value is below 0.5.

The Receiver Operating Characteristics (ROC) curve of LRM demonstrates ability of classification based on the variation of cutoff values. The construction of ROC curve at various cutoff values is plotted by True Positive Rate (TPR)

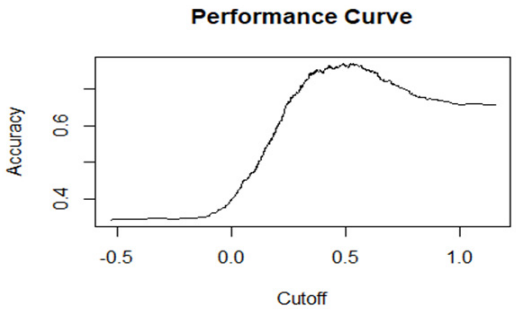

FIGURE 5. Training data's Performance curve.

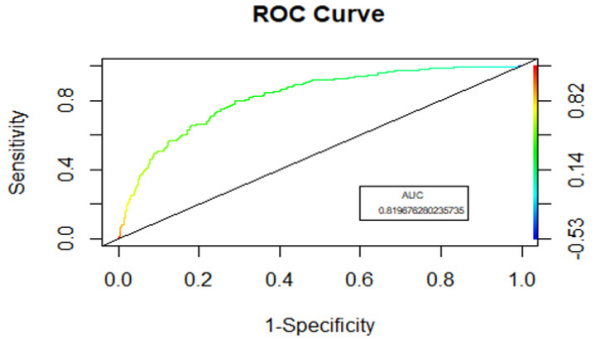

FIGURE 6. Training data's ROC curve.

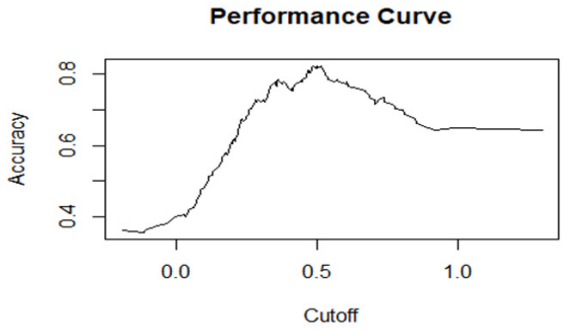

FIGURE 7. Testing data's Performance curve.

against False Positive Rate (FPR). Here TPR is named as Sensitivity and calculated FPR is known as (1- Specificity). Hence ROC curve is used to analyze the optimality of LRM. The ROC curve is interpreted by calculating value of Area Under Curve (AUC). The overall ROC curve value is 1 which is the cumulative probability distribution value. A slope is drawn and used as a partition to analyze a ROC curve. The ROC curve if present above the slope indicates that LRM prediction is more than 50% where as if ROC curve if present below the slope then LRM prediction is less than 50% which is not good for a model.

The snapshot of ROC curve of training data of PIDD is shown in Figure 6. In this Figure 6 the prediction ROC curve is distributed over various cutoff values ranging between −0.53 and 0.82. The LRM prediction of ROC curve is more than 50% as the curve is above the slope. It is observed from

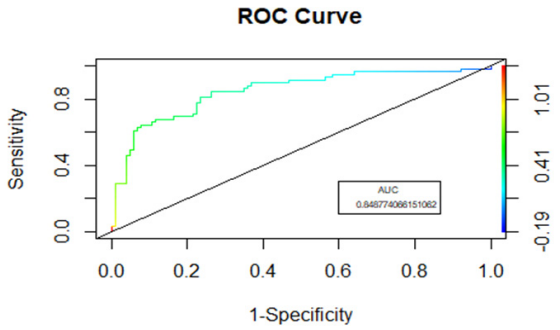

FIGURE 8. Testing data's ROC curve.

TABLE 3. Comparison of outcomes of PIDD with several existing methods.

Sl. No.	Previous work by	Method	Accuracy
1	Polat et al. [1]	GDA–LS-SVM (10x FC)	79.16%
2	Seera Manjeevan et al.[3]	FMM-CART-RF	78.39%
3	Dilip Kumar Choubey et al. [6]	GA-NBs	78.6957%
4	S.Sa'di et al. [4]	NB's	76.95%
5	Aakanksha Mahajan et al. [5]	PSO-knn	77%
6	Sisodia, Deepti et al. [11]	NB's	76.30%
7	Ahmad et al. [12]	MLP	81.9%
8	Our study	PCA-LRM	82.1%

the Figure 6 that near 0.5 cutoff value, TPR value is high. The calculated AUC value of training data is 0.8196.

From PIDD, 20% of data is considered for testing the data. The snapshot of performance curve of testing data of PIDD is shown in Figure 7. The performance curve of LRM is analyzed with respect to accuracy and cutoff value used in disease prediction. Figure 7 depicts that at cutoff value 0.5, LRM prediction accuracy raises above 80%. LRM prediction accuracy falls low when the cutoff value is below 0.5.

The snapshot of ROC curve of testing data of PIDD is shown in Figure 8. In this Figure 8 the prediction ROC curve is distributed over various cutoff values ranging between −0.19 and 1.01. The LRM prediction of ROC curve is more than 50% as the curve is above the slope. It is observed from the Figure 8 that near 0.5 cutoff value, TPR value is high. The calculated AUC value of testing data is 0.8487.

There are various existing, implemented methods on PIDD. Few are mentioned in Table 3, which gives details about the performance measure of all the executed methods. It is analyzed from Table 3 that PCA-LRM achieves improved accuracy on PIDD than other prevailing methods.

V. CONCLUSION AND FUTURE WORK

Feature extraction and statistical modeling on PIDD is presented in this research work. The PIDD features are extracted to a new space using PCA. These newly projected features are then modeled using LRM to predict whether the patient is diabetic or normal. The results obtained in this study have achieved high accuracy rate for predicting diabetes when compared with other existing methods. The proposed statistical model can be adopted for predicting different kinds of diseases like tuberculosis, eye disease, cancers, etc.,

REFERENCES

[1] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 482–487, 2008.
[2] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14650–14659, 2011.
[3] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, 2014.
[4] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, "Comparison of data mining algorithms in the diagnosis of type II diabetes," *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 1–12, 2015.
[5] R. Bansal, S. Kumar, and A. Mahajan, "Diagnosis of diabetes mellitus using PSO and KNN classifier," in *Proc. Int. Conf. Comput. Commun. Technol. Smart Nation (IC3TSN)*, Oct. 2017, pp. 32–38.
[6] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection," in *Proc. Int. Conf. Commun. Comput. Syst. (ICCCS)*, Feb. 2017, pp. 451–455.
[7] H. Roopa and T. Asha, "Feature extraction of chest X-ray images and analysis using PCA and kPCA," *Int. J. Elect. Comput. Eng.*, vol. 8, no. 5, p. 3392, Oct. 2018.
[8] H. Roopa and T. Asha, "Analysis of feature ranking methods on X-ray images," in *Proc. Int. Conf. ISMAC Comput. Vis. Bio-Eng.* Cham, Switzerland: Springer, 2018, pp. 1393–1403.
[9] *UCI Repository of Machine Learning*. Accessed: 2017. [Online]. Available: http://www.ics.uci.edu./~mlearn/ML Repository.html
[10] H. Roopa and T. Asha, "An efficient prediction model to analyze tuberculosis chest X-ray images," *Int. J. Sci. Eng. Res.*, vol. 10, no. 1, pp. 694–700, 2019.
[11] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, Jan. 2018.
[12] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, "Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus," in *Proc. Int. Conf. Digit. Inf. Process. Commun.* Berlin, Germany: Springer, 2011, pp. 537–545.