A Hybrid Data Mining Approach for Diabetes Prediction and Classification

Vemuri Bharath Kumar, Kumba Vijayalakshmi, *Member, IAENG* and M. Padmavathamma

Abstract— **Diabetes has emerged as the most chronic disease that may cause mortality to the diabetic patients. However, early prediction of diabetes can be helpful in reducing the severe effects of diabetes. Several approaches have been developed for diabetes prediction, recently, the data mining based machine learning approaches have gained huge attraction from research community. However, conventional approaches suffer from the performance related issues. Hence, in this work we introduce a novel data mining approach for diabetes prediction and classification. The proposed approach includes missing value imputation, data clustering, dimension reduction and Bayesian regularized neural network classification. The proposed approach is carried out using open source available Pima Indian Diabetes dataset and implemented using MATLAB simulation tool. The obtained performance is compared with the existing techniques; this comparative study shows a significant improvement in the prediction performance using proposed approach.**

Index Terms—**Bayesian regularized neural network classification, Diabetes, Data clustering, Dimension reduction machine learning data clustering, dimension reduction**

I. Introduction

DIABETES Mellitus (DM) is considered as one of the most chronic and deadly disease because it can cause serious threats to the human health. Recently, a study revealed that total 552 million people are expected to get affected due to DM by 2030 [1]. Generally, diabetes is classified into three main subtypes as Type 1 diabetes, Type 2 diabetes and gestation diabetes. The Type 1 diabetes is characterized, due to the deterioration of the beta cells which are responsible for insulin production in the human body, similarly, Type 2 diabetes is characterized by insulin secretion and insulin resistance in the human body. Gestation diabetes affects the pregnant woman. According to the study presented in [5], approximately 4.9 million deaths has occurred due to the diabetes related issues.

The late diagnosis of diabetes raises more complexities in the treatment of diabetes. Barakat et. al. [6] suggested that the early identification of these diseases can help to prevent and delay the total 80% of complication of Type 2 diabetes. Several techniques have been reported for diabetes prediction such bio-medical signal processing which includes EMG (Electromyography) signals [7], computer vision techniques such as diabetic retinopathy [8], and machine learning techniques such as data mining [9].

Data mining based techniques have been adopted in various medical related applications. Asri et al.[10] discussed the advantages of data mining for medical application and presented a comparative study for breast cancer risk prediction using different classification schemes such as Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k-Nearest Neighbors (k-NN) techniques. The breast cancers are classified as benign and malignant according to their severity. Chaurasia et. al. [11] presented a data mining approach for predicting the benign and malignant stages using data mining approach by using Naïve Bayes, RBFNetwork, and J48 classification schemes. Similarly, in [12] authors presented genetic algorithm based feature selection scheme for breast cancer prediction where different types of classifier such as Logistic Regression, Decision Trees Random Forest, Bayesian Network, Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN) and Support Vector Machine (SVM) are used.Kannan et. al. [13] discussed about the heart disease related issues and presented a comparative analysis of different types of data mining based approach for heart disease prediction. Similarly, Long et al. [14] presented optimization scheme based approach for heart disease prediction. In this work, roughest based feature reduction and interval-Type-2 fuzzy logic (IT2FL) approach is developed. The Type-2 fuzzy logic is combined with the fuzzy-c-means clustering, firefly optimization and genetic algorithm. Similarly, Paulet. al. [15] developed fuzzy rule based heart disease prediction. In this work also, the fuzzy logic model is combined with the modified dynamic multi-swarm particle swarm optimization (MDMS-PSO) to obtain the optimized attribute set.

On the other hand, data mining based approaches are also used for diabetes prediction. Anto et al. [16] presented least square support vector machine for diabetes prediction. In this work, the SVM and simulated annealing are combined together for improve the prediction accuracy. Moreover, this work includes Fisher score (FS) for selecting the most significant attributes. According to this approach, the least square SVM and RBF (radial basis function) classifier are applied to perform the classification and simulated annealing is applied for optimizing the kernel parameters of LS-SVM.

Numerous techniques are present in this field where selection of appropriate machine learning model is a challenging task. Recently, Nilashiet al. [17] presented a new technique to improve the classification accuracy. In this work, authors developed data clustering based scheme for

Manuscript received May 16, 2019; revised August 05, 2019.

Vemuri Bharath Kumar is pursing Bachelor degree in Computer Science Department at Sri Venkateswara Vedic University, Tirupati, AP 517502 India (e-mail: bharath.rojasudharshan@gmail.com).

Kumba Vijaylakshmi is with the Department of Computer Science as Assistant Professor, Sri Venkateswara University, Tirupati, AP 517502 India. (e-mail: vijayalakshmik4@gmail.com).

M. Padmavathamma is with the Department of Computer Science as Professor, Sri Venkateswara University, Tirupati, AP 517502 India.

classification of diabetes. This work mainly can be divide into three stages where first of all, SOM clustering is applied to group the various data based on their attributes, later PCA scheme is presented for dimensionality reduction and finally, neural network classifier is applied to obtain the classification performance. Mercaldo et al. [18] used HoeffdingTree algorithm for diabetes prediction. Similarly, Sangleet. al. [19] presented a diabetes detection model using a combination of principal component analysis and Multilayer Perceptron Artificial Neural Network (MLPANN). Nilashi et al. [20] presented soft computing based approach where data clustering, noise removal from the data and classification stages are developed. In order to achieve the desired tasks, expectation maximization is applied for clustering; PCA for noise removal and classification is performed using SVM classifier.

II. Proposed model

A. Review Stage

In this section we present the proposed solution for diabetes prediction using data mining based approach. the proposed model is divided into four main stage which are as follows: (a) pre-processing, (b) clustering (c) dimension reduction and (d) classification. According to the first phase of proposed approach, we perform the data pre-processing. In this work, we have considered UCI repository diabetes dataset known as Pima Indian Diabetes Dataset [21] which contains missing attributes thus missing value imputation is a main phase to achieve the better accuracy. Several techniques have been reported for missing value imputation as described in [22, 23] but the medical data is collected from the new observations hence conventional approaches may suffer from the poor performance of data imputation. Thus, the first objective of this work is to develop an adaptive process for missing value imputation. In the next phase, we apply, data clustering strategy which helps to group the diabetes attributes based on their similarities. These cluster groups reduces the misclassification error during database training and testing. Several clustering approaches have been reported for dimension reduction and cluster formation such as [24, 25] but there exist some challenges with these clustering algorithms such as selection of distance computation algorithm for categorical attributes, selection of appropriate cluster numbers, types of attributes and selection of initial cluster. Due to these issues, the clustering performance degrades hence our second objective is to present a clustering novel clustering algorithm with less clustering error. In the third phase, attribute selection is considered as the key component which has a noteworthy impact on the prediction and classification using data mining techniques. The proper attribute selection helps to select the attributes which are most suitable to achieve the high accuracy due to the most relevance features. Based on this assumption, several techniques are presented for feature selection and dimension reduction such as Fuzzy Logic [23], and PCA based approach [26] etc. Moreover, these schemes are divided as linear (some well-known techniques such as Principal Component Analysis, Linear Discriminate Analysis) and non-linear methods (some well-known methods are Isometric Feature Mapping and Locally Linear Embedding etc.) of attribute selection but the existing approaches suffer from several challenging issues such as scalability because of the increasing dimensionality new search mechanisms are required. Computational complexity is also considered as a challenging task because the existing methods require large scale computation process to solve the dimension reduction issue, moreover, these approaches require more time hence a new approach for dimension reduction is recommended which can reduce the complexity for development of a better model of data mining. Finally, we present classification model where the selected attributes are trained and tested to obtain the classification performance of proposed model. Ndaba et. al. [27] presented regression neural network based classification model for diabetes prediction using Pima Indian Diabetes dataset. Wu et. al. [28] developed K-means and logistic regression algorithm for predicting type 2 diabetes mellitus. Similarly, Mohapatraet al. [29] introduced Deep Neural Network (DNN) based approach for diabetes detection. Tan et. al. [29] presented a combined approach using genetic and fuzzy logic algorithm for classifying the diabetes data patterns. In this field of data mining in diabetes classification, ensemble classifiers have gained attraction due to their significant learning process which helps to improve the accuracy. Based on this assumption, Garg et al. [30] developed an ensemble classifier where neural network classifier and decision tree algorithms are combined together and a binary classifier is developed. Similarly, Kumar et al. [31] also developed an ensemble classifier by using Random Forest and Gradient Boosting classifiers models.

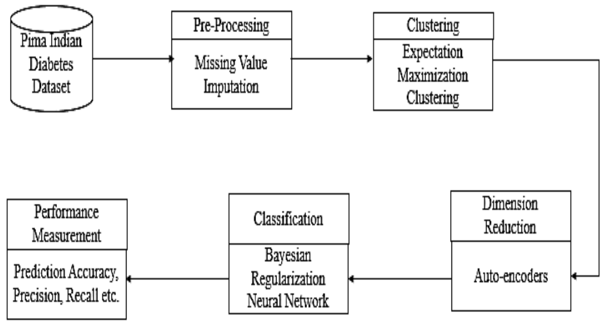
Pima Indian Diabetes Dataset
Pre-Processing
Missing Value Imputation
Clustering
Expectation Maximization Clustering
Dimension Reduction
Auto-encoders
Classification
Bayesian Regularization Neural Network
Performance Measurement
Prediction Accuracy, Precision, Recall etc.

Figure 1: Proposed Methodology for prediction of Diabetes

Several approaches are presented but the accuracy of diabetes prediction rely on the all phases of data mining as discussed before. Moreover, classifier training models suffer from the learning error which causes misclassification during misclassification and degrades the system performance. Hence, a novel algorithm is required for learning the diabetes data patterns efficiently. In this work, we focus on these challenges of diabetes prediction using data mining and develop a new approach for diabetes prediction and classification. Figure 1 show the flow of proposed approach where data pre-processing, data clustering, dimensionality reduction and classification techniques are performed. Based on this classification, finally, performance measurement parameters are computed and obtained.

B. *Missing value imputation*

In this section, we present the proposed solution for missing value imputation to improve the accuracy of classification. Generally, K-NN based approach achieves better performance for missing value identification and imputation but due to computational complexities, the KNN approach fails to provide the notable performance for large scale datasets. Hence, in this work we present the weighted adaptive data imputation scheme for data mining. This approach of missing value imputation is known as weighted Adaptive Data Imputation scheme (WAIDS).

The proposed approach can be performed for each class of the considered diabetes data for training. In this process, we form a weighted vector as $P \times Q$ which represents the topological features as weighted vector of the entire class and also these feature represents the corresponding classes. In this process, we create small size weighting vectors which helps to obtain the neighboring clusters with less computational complexity. At this stage, the obtained weighted vector v for class $w_c$where $= 1,2, \dots C,$ is expressed as $\alpha_v^{w_c}, v = 1,2, \dots . V$.

In order to estimate the missing values, selected weights choose close weighting vectors and the weight $p_{iv}^{w_c}$of each vector is computed based on the distance between current object x and obtained weighting vector as $p_{iv}^{w_c} = e^{\left(-\lambda d_{i_v}^{w_c}\right)}$ where $\lambda = \frac{cQP(cQP-1)}{2\sum_{i,j} d(\alpha_i,\alpha_j)}$ and $d_{i_v}^{w_c}$ denotes the Euclidean distance between x and neighboring vector and $d\left(\alpha_i, \alpha_j\right)$ represents the distance between any two weighting vectors. Here, the selected weighting vector for class w_c are used for computing the weighted mean as $\hat{y}_i^{w_c}$, this weighted mean is used as imputing the missing values and it can be obtained as:

$$\hat{y}_i^{w_c} = \frac{\sum_{v=1}^{V} p_{iv}^{w_c} . \alpha_v^{w_c}}{\sum_{v=1}^{V} p_{iv}^{w_c}} \qquad (1)$$

The obtained missing values are imputed in the data in the same dimensions and the final pre-processed data can be obtained for further process.

TABLE I
NOTATIONS USED

Notation	Description
v	Weighted data vector
w	Current class
C	Total available class
α	Selected weighted vector
p	Weight of current vector
x	Data
z	Missing data components
$\mathcal{L}$	Log-likelihood
$\mathbb{D}$	Dataset
$\mathbb{N}$	Neural Network Model

C. *Data clustering*

This section presents the data clustering model for diabetes class prediction. The clustering process helps to group the similar attributes according to the diabetes group. This process also helps to predict the diabetes even if the class labels are not present or data is unsupervised. Hence, it increases robustness of the classification performance. In this work, we use Expectation Maximization approach for data clustering. According to the EM process, the data is represented as $(x.z).z$ where z represents the missing data components from the labels of observation denoted as $z = (z_1, \dots z_n)$where $z_i = k$, if x_i belongs to the k component. The complete log-likelihood can be expressed as:

$$\mathcal{L}(\varphi, Z|X) = \sum_{k=1}^{K}\sum_{k=1}^{K} z_{ik} \log\left(\pi_k f_k(x;\theta_k)\right) \qquad (2)$$

In this process, the EM approach initiates with the initial parameter as θ^0, later, this process is divided into two computation phases as E step computation and M step computation.

According to the E step, the expected likelihood value is computed with respect to the condition distribution of z for the given x under the estimate parameter φ,this can be computed as:

$$Q\left(\varphi, \varphi^{(q)}\right) = E[P]$$
$$P = \log\left(CL(\varphi, Z|X)\right) \qquad (3)$$

This is used for computing the posterior probabilities of a data to belong to the k^{th} component as:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} f_k(x;\theta_k^{(q)}}{\sum_l \pi_l^{(q)} f_l\,(x;\theta_l^{(q)}} \qquad (4)$$

In the next stage, we perform the M step where φ^{q+1} is considered which maximizes the current expectation. It is given as:

$$\varphi^{q+1} = \underset{\varphi}{argmax}\; Q\left(\varphi|\varphi^{(q)}\right) \qquad (5)$$

D. *Dimensionality reduction*

As discussed in previous section we have described that dimensionality reduction methods are divided as linear and non-linear methods but these methods suffer from various issues hence we present auto encoder based dimensionality reduction method. Let us consider that the input data x contains n dimensional subspace which needs to be decomposed into y belonging to the m dimensional data as reduced dimensions. In order to obtain the dimensionality reduction, we consider the training of auto encoder. Generally, the encoders and decoders are utilized during training, but after finishing the training, the encoder is used and decoder is discarded from the process. Figure 2 shows the architecture of auto encoder where inputs and the outputs of the neural network are presented.

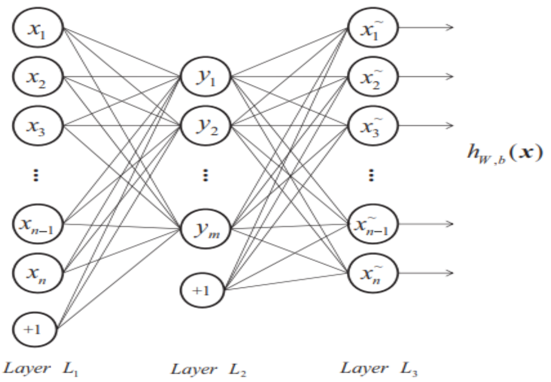

Figure 2 : Architecture of auto encoder

Hence, in order to obtain the dimensionality reduction, a low dimensional layer needs to be incorporated between encoder and decoder layer. Let us consider a three-layered neural network where output of the neural network is $h_{w,b}\,(x) = (\bar{x}_1, \bar{x}_2, .., \bar{x}_n)^T$ equal to the input $x = (x_1, x_2, \dots x_n)^T$ and J denotes the reconstruction error, here we present the back propagation algorithm for training the neural network. This is given as:

$$h_{w,b}\,(x) = g\big(f(x)\big) \approx x$$
$$J(W, b; x, y) = \frac{1}{2}\left\|h_{W,b}(x) - y\right\|^2 \qquad (6)$$

Generally, auto encoder can be considered as a method for transform representation. In this process, if the number of hidden layer is greater than the original inputs along with the sparsity constraint, the network behaves as sparse coding. Similarly, the if the hidden layer nodes are less than the original input then it gives the compressed representation of the considered inputs which is known as the reduced dimension of the data.

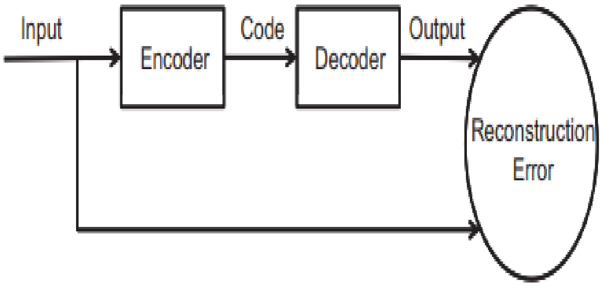

Figure 3 :Dimension reduction using auto encoder

Figure 3 shows a process of dimension reduction using auto encoder where the first layer to second layer f data is processed and second layer to third layer data is processed to decoder and reconstruction error is minimized by adjusting the parameters of encoder and decoder.

E. Diabetes prediction and classification

In this section, we present the classifier model for diabetes classification using Bayesian regularization neural network. In this field of NN based learning, the back propagation neural network is considered as a most promising technique which uses gradient descent method to reduce the error in terms of mean squared error (MSE). Neural network contains three main layers which are known as input layer, hidden layer and output layer. This process follows supervised learning scheme weights are adjusted according to the outcome to reduce the error which helps to optimize the network performance for unknown samples. However, these techniques suffer from the issue of overtraining and overfitting.

In order to mitigate these issues, we utilize Bayesian regularization model which helps to convert the non-linear problem to a structured problems. Here our main aim is to reduce the training error. According to the Bayesian regularization method, the BP network can be expressed as:

$$\mathcal{F} = \beta\mathcal{E}_D + \alpha\mathcal{E}_w \qquad (7)$$

Where $\mathcal{E}_w = \frac{1}{2}\sum_{i=1}^{N} w_i^2$ which is the sum of network weights and $\mathcal{E}_D$ is the sum of squared errors, α and β represents the objective functions of regularization parameters. According to the Bayesian network, the network weights are considered in the form of random variables which can be updated using Bayes rule, this can be given as:

$$P(w|D, \alpha, \beta, M) = \frac{P(D|w, \beta, M)P(w|\alpha, M)}{P(\mathbb{D}|\alpha, \beta, \mathbb{N})} \qquad (8)$$

Where $\mathbb{D}$ denotes the dataset, $\mathbb{N}$ denotes the neural network model and w represents the weight vector. Here probability density function is computed for each data to estimate the occurrence of particular data. This approach reduces the local minima problem and increases network generalizability.

Similarly, we present a neural network model where input weights are optimized according to the mean squared error. Here, we consider a tangent sigmoid function which is computed as:

$$\tanh(n) = \frac{2}{1 + \exp(-2n)} - 1 \qquad (9)$$

This model gives us the predicted output based on the trained diabetes data.

III. Results and Discussion

In this section, we present the experimental study and performance comparison of proposed approach of diabetes prediction with other classification techniques. The proposed approach is implemented using MATLAB simulation tool running on windows platform with i5 processor and 8GB RAM. In order to evaluate the performance in terms of accuracy, precision, recall, sensitivity, specificity, and F-measure. A brief discussion about these measures is presented in below given section.

A. Performance measurement metrics

Accuracy is a measurement of rate of correct classification. It can be computed by taking the ratio of correct classification and total number of instances. It can be expressed as:

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \qquad (10)$$

Another performance measurement parameter is known as sensitivity, which is computed as given in (11)

$$Sensitivity = \frac{T_P}{T_P + F_N} \qquad (11)$$

Next parameter is computed as given in (12):

$$Specificity = \frac{T_N}{TN + FP} \qquad (12)$$

Then, we compute Precision of the proposed approach. It is computed by taking the ratio of True Positive and (True and False) positives.

$$P = \frac{TP}{TP + FP} \qquad (13)$$

Finally, F-measure is computed which is the mean of precision and sensitivity performance. It is expressed as:

$$F = \frac{2 * P * S_N}{P + S_N} \qquad (14)$$

B. Database details

The proposed approach is implemented and experiments are conducted in Pima Indian Diabetes database which are obtained from UCI repository. A brief detail of these dataset is given in this section.

The considered PID dataset contains total 768 instances which are obtained from different patients, each data contains total 768 attributes. The diabetic and non-diabetic classes are indicated as “0” and “1” which reports the negative and positive test for considered patient. The clinical attributes are represented as:

- Number of times pregnant (NP).
- Plasma glucose concentration after 2 h in an OGTT.
- Diastolic blood pressure (mmHg) (DBP).
- Triceps skinfold thickness (mm) (TSFT).
- Two-hour serum insulin (μU/mL) (2HSI).
- BMI.
- Diabetes pedigree function (DPF).
- Age (years) (AGE).

This dataset contains total 500 sample as non-diabetic class and 268 instances are collected from diabetic class. Based on these parameters we evaluate the prediction accuracy of the proposed model.

C. Comparative study

In this section we present the performance measurement and comparative study of proposed model with existing approaches of data mining based machine learning.

In this work we have considered Bayesian regularized neural network where we have used 70% data for training purpose i.e. 538 samples for training, 15% data for data validation and testing i.e. 115 data samples for testing and validation, respectively. In this network, total 10 number of hidden layers are considered. Based on this configuration, the obtained neural network architecture is obtained as depicted in figure 4.

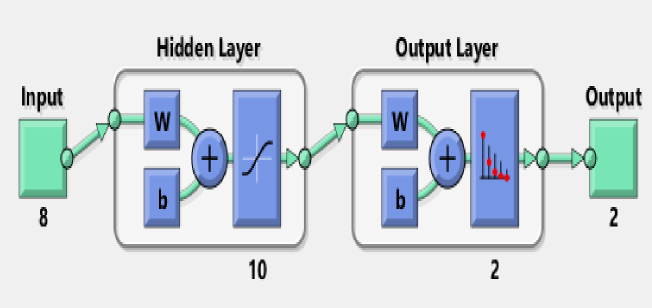

Figure 4 :Neural Network Architecture

Similarly, we apply feature selection scheme for dimensionality reduction and the obtained neural network architecture is presented in figure 5.

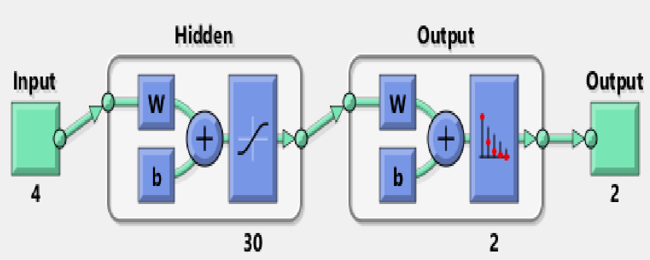

Figure 5: Neural Network Architecture after dimension reduction

With the help of this technique, we obtain the confusion matrix as presented in figure 6. According to this figure the overall accuracy of the system is obtained as 94.5%.

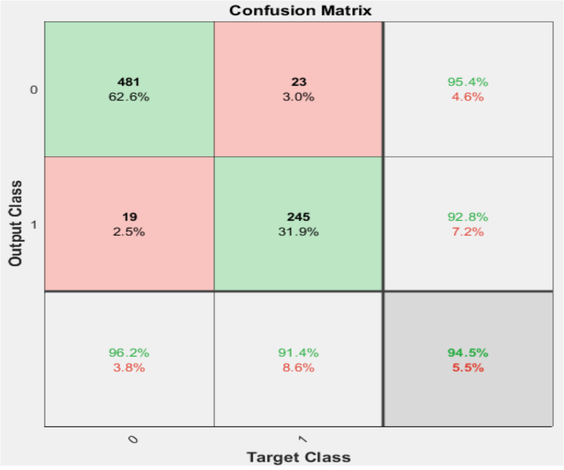

Figure 6 : Confusion matrix plot

Furthermore the precision is obtained as 91.41%, sensitivity is obtained as 92.80%, and specificity is obtained as 95.43%.

Based on the obtained performance, we compare the performance of proposed approach with existing techniques as presented in Table 2.

TABLE II
PREDICTION ACCURACY COMPARISON

Technique	Accuracy %
Kayaer et. al. [33]	80.21%
Polat et. al. [34]	79.16%
Çalişir et. al. [35]	89.74%
Erkaymaz et. al. [36]	91.66%
HPM [25]	92.38%
Marcano-Cedeño et al. [38]	89.93%
Proposed approach	**94.5%**

The comparative study shows that the proposed approach achieves better accuracy for diabetes prediction when compared with the existing techniques.

IV. Conclusion

In this work, we have focused on the data mining technique to predict and classify the diabetes. The proposed approach is a combination of several steps of data mining. Initially, we apply missing value imputation and developed a weighed adaptive approach for computing and identifying the missing values. In the next phase, we present EM-clustering algorithm, later, dimension reduction approach is applied using auto-encoders and finally Bayesian regularized classifier is presented. The performance of proposed approach is obtained as 94.5% in terms of prediction accuracy.

References

[1] Whiting DR, Guariguata L, Weil C, et al. IDF diabetes atlas: "Global estimates of the prevalence of diabetes for 2011 and 2030, Diabetes" Res. Clin. Pract. , 2011, vol. 94 (pg. 311-321).

[2] Chang K-H, Chuang T-J, Lyu R-K, et al." Identification of gene networks and pathways associated with Guillain-Barre syndrome" PloS One , 2012, vol. 7 (pg. 80-89).

[3] Ramachandran A, Snehalatha C, Kapur A et al (2001) "High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Surve". Diabetologia 44:1094– 1101.

[4] Ramachandran, A. (2004). "Diabetes & obesity-the Indian angle" Indian Journal of Medical Research, 120(5), 437.

[5] Indian Dental Association 6th Edition Committee: "Follow-up to the political declaration of the high-level meeting of the General Assembly on the prevention and control of non-communicable diseases" 6th edn. IDF (2013).

[6] Barakat, N.H., Bradley, A.P., Barakat, M.N.H.:" Intelligible support vector machines for diagnosis of diabetes mellitus" Trans. Info. Tech. Biomed. 14(4), 1114–1120 (2010).

[7] Caesarendra, W., Lekson, S. U., Mustaqim, K. A., Winoto, A. R., & Widyotriatmo, A. (2016)," A classification method of hand EMG signals based on principal component analysis and artificial neural network", 2016 International Conference on Instrumentation, Control and Automation (ICA). doi:10.1109/ica.2016.7811469.

[8] Gargeya, R., & Leng, T. (2017) " Automated identification of diabetic retinopathy using deep learning", Ophthalmology, 124(7), 962-969.

[9] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018),
" Predicting diabetes mellitus with machine learning techniques" Frontiers in genetics, 9.

[10] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016),
" Using machine learning algorithms for breast cancer risk prediction and diagnosis" Procedia Computer Science, 83, 1064-1069.

[11] Chaurasia, V., Pal, S., & Tiwari, B. B. (2018)," Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology, 12(2), 119-126.

[12] Aličković, E., & Subasi, A. (2017),"Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications", 28(4), 753-763.

[13] Kannan, R., & Vasanthi, V. (2019),"Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease" in Soft Computing and Medical Bioinformatics (pp. 63-72). Springer, Singapore.

[14] Long, N. C., Meesad, P., & Unger, H. (2015) "A highly accurate firefly based algorithm for heart disease prediction" Expert Systems with Applications, 42(21), 8221-8231.

[15] Paul, A. K., Shill, P. C., Rabin, M. R. I., & Murase, K. (2018). "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease", Applied Intelligence, 48(7), 1739-1756.

[16] Anto, S., Chandramathi, S., & Aishwarya, S. (2016),"An expert system based on LS-SVM and simulated annealing for the diagnosis of diabetes disease", International Journal of Information and Communication Technology, 9(1), 88-100.

[17] Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L. (2017)," Accuracy improvement for diabetes disease classification: a case on a public medical dataset" Fuzzy Information and Engineering, 9(3), 345-357.

[18] Mercaldo, F., Nardone, V., & Santone, A. (2017)" Diabetes mellitus affected patients classification and diagnosis through machine learning techniques", Procedia computer science, 112, 2519-2528.

[19] Sangle, S., Kachare, P., & Sonawane, J. (2019)," PCA Fusion for ANN-Based Diabetes Diagnostic" in Computing, Communication and Signal Processing (pp. 583-590). Springer, Singapore.

[20] Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A., & Jusoh, A. (2018)," A soft computing approach for diabetes disease classification" Health Informatics Journal, 24(4), 379-393.

[21] Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/ machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes. data. Accessed: 1st of May, 2008.

[22] Purwar, A., & Singh, S. K. (2015)," Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications" 42(13), 5621-5631.

[23] Dzulkalnine, M. F., & Sallehuddin, R. (2019)" Missing data imputation with fuzzy feature selection for diabetes dataset", SN Applied Sciences, 1(4), 362.

[24] Santhanam, T., & Padmavathi, M. S. (2015).,"Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis", Procedia Computer Science, 47, 76-83.

[25] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010)," Hybrid prediction model for type-2 diabetic patient" Expert systems with applications, 37(12), 8102-8108.

[26] Kale, A. P., & Sonavane, S. (2018). PF-FELM: "A Robust PCA Feature Selection for Fuzzy Extreme Learning Machine" IEEE Journal of Selected Topics in Signal Processing, 12(6), 1303-1312.

[27] Ndaba, M., Pillay, A. W., & Ezugwu, A. E. (2018, May)." An improved generalized regression neural network for type II diabetes classification" in International Conference on Computational Science and Its Applications (pp. 659-671). Springer, Cham.

[28] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). "Type 2 diabetes mellitus prediction model based on data mining" Informatics in Medicine Unlocked, 10, 100-107.

[29] Mohapatra, S. K., Nanda, S., & Mohanty, M. N. (2018, April). "Diabetes Detection Using Deep Neural Network", In International Conference on Soft Computing Systems (pp. 225-231). Springer, Singapore.

[30] Tan, C. H., Tan, M. S., Chang, S. W., Yap, K. S., Yap, H. J., & Wong, S. Y. (2018). "Genetic algorithm fuzzy logic for medical knowledge-based pattern classification" Journal of Engineering Science and Technology, 13, 242-258.

[31] Garg, D., & Mishra, A. (2018). "Bayesian regularized neural network decision tree ensemble model for genomic data classification" Applied Artificial Intelligence, 32(5), 463-476.

[32] Kumar Das, S., Kumar Mishra, A., & Roy, P. (2019). "Automatic Diabetes Prediction Using Tree Based Ensemble Learners" International Journal of Computational Intelligence & IoT, 2(2).

[33] Kayaer K and Yıldırım T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: Proceedings of the international conference on artificial neural networks and neural information processing, 26–29 June 2003, pp. 181–184. Istanbul: Springer.

[34] Polat K, Günes S and 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): 482–487

[35] Çalişir D and Doğantekin E. "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier" Expert Syst Appl 2011; 38(7): 8311–8315.

[36] Erkaymaz O and Ozer M. "Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes" Chaos Soliton Fract 2016; 83: 178–185.

[37] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010)." Hybrid prediction model for Type-2 diabetic patients" Expert Systems with Applications, 37(12), 8102–8108. doi:10.1016/j.eswa.2010.05.078.

[38] Marcano-Cedeño, A., Torres, J., & Andina, D. (2011, May). "A prediction model to diabetes using artificial metaplasticity" In International Conference on the Interplay Between Natural and Artificial Computation (pp. 418-425). Springer, Berlin, Heidelberg.