Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 6, June 2020, pg.132 – 140

GENERIC FRAMEWORKS FOR SVM, ANN, LGBM, AND LR ALGORITHMS

Nora Ibrahem Alghurair[1]; Mohammad A. Mezher[2]
[1]College of Postgraduate Studies & Scientific Research, Fahad Bin Sultan University, Tabuk, KSA
[2]College of Computing, Fahad Bin Sultan University, Tabuk, KSA
[1] noraibrahem018@gmail.com; [2] mmezher@fbsu.edu.sa

Abstract*— World Health Organization describes diabetes as a multiple etiological metabolic condition defined by persistent hyperglycemia with anomalies in glucose, lipid and protein metabolism triggered by insulin secretion deficiencies, lipid or both. Diabetes is one of the 21st century's most daunting health problems in the world, and affects over 425 million people. Data mining is one of the major techniques which develops and supports medical data research. The aim of this research was to establish a diagnostic framework for diabetes. The PIDD used to function description. Two generic frameworks have been proposed in the study. The first framework uses the ANN technique and fed production to SVM which yields the diagnostic result. With this generic framework, five experiments are carried out and the highest accuracy achieved was 81.8%. The second framework employs an ensemble of majority voting techniques which combines LGBM, SVM, and LR. The generic framework was right at 87.9 per cent. The frameworks presented in contrast with other state-of-the-art solutions, and it found that the second solution is the better one.
Keywords*— Neural Networks, LGBM, SVM, LR, Diabetes, PIMA Indian, UCI

I. INTRODUCTION

The World Health Organization (WHO) describes diabetes as a multiple etiological metabolic condition defined by persistent hyperglycemia with glucose, lipid and protein metabolism abnormalities triggered by deficiencies in insulin secretion, lipid, or both (Lorenzo et al.,2007). Nearly half of the patients with diabetes was affected by heredity factor, which is known to be a significant characteristic of diabetes. The pathologic causes of diabetes are the decreased insulin output leading to pancreatic failure. There are two forms of Diabetes. β-cells are type 1 diabetes mellitus pathogenesis, which is impaired by pancreatic secretions, blocking the lowering of blood glucose level. The impairment of insulin resistance and insulin secretion, the pathogeneses of type 2 diabetes mellitus are known to be diabetes based on non-insulin (Muller et al.,2005).

In under fifty years, the incidence of this disorder has risen five-fold. This gradual increase is due to different factors. World population growth, a rise in life expectancy for diabetics, a rise in pregnancy among diabetic mothers, an increase in obesity, an increase in intake of processed sugars. Along with other factors that can serve as a trigger such as sedentary lifestyle, diets high in saturated fat and proteins, reduced fiber intake, a food poor in complex carbohydrates and vitamin E, unnecessary stress, Tobacco that may promote the resistance to insulin (Vlassara & Uribarri.,2014).

Diabetes is one of the world's most daunting health problems of the 21st century and impacts more than 425 million individuals. The most prevalent type 2 diabetes mellitus triggered by insulin resistance is the most severe form of diabetes accounting for 90-95 cases led by type 1 diabetes mellitus accounting for diabetes becoming the world's fifth leading cause of disease death (Wild et al.,2004).

Diabetes is becoming a significant global problem that is viewed as a red threat to human health. Diabetes is linked with serious symptoms and a variety of adverse reactions of all sorts including lack of hearing, impaired immune function, cardiac problems, skin disorders, renal dysfunction, nerve impairment and even disruption to

the blood vessels. There is no full remedy and diabetes care which will absolutely cure the diabetes (Gill & Mittal., 2016). So, detecting the illness at an early stage is important to mitigate its harmful impact on human health. However, it is nearly difficult to totally cure the diabetes, but it can be regulated enough that an individual can live a healthier life. Late disease detection is critical and is one of the key thrust fields of the research (Gill & Mittal., 2016). A practical intelligent method will be utilized to help the doctors to create an early and effective diagnosis.

Given that there are growing increases in the healthcare sector, it is particularly beneficial to ensure patients take custody of diabetics on their own. For certain situations, the original knowledge relating to diabetics assists with avoiding diabetes, healing it and managing the condition in an appropriate way. Many computer systems were developed by embedding the human intelligence that is helpful in handling diabetes for the victims (World Health Organization,. 2020).

In evaluating different technologies, such as artificial intelligence, cell phone apps and tools expressly built for diagnosing diabetic disease.

Data mining is one big technique that develops and assists the area of medical data research. In the field of healthcare, data mining is becoming more widespread because in clinical data there is a need for effective methods to find unknown and useful knowledge. This also lets healthcare researchers design healthcare strategies that are successful, create medication prescription programs; establish patient health profiles, etc. Clinical data-mining entails conceptualizing, collecting, reviewing and evaluating accessible clinical evidence for knowledge-building, professional decision-making and reflection by practitioners (Deepika & Poonkuzhali., 2015).

Through this research, the goal is to establish a diabetes diagnostic framework. The Pima Indians Diabetes Database of National Institute (PIDD) of Diabetes and Digestive and Kidney Diseases is used for classification task. In the analysis two methods were suggested. The first methodology employs Artificial Neural Network (ANN) strategy and feed results to Support Vector Machine (SVM) that outputs the diagnostic. Five experiments are carried out with this framework and the highest accuracy achieved was 81.8%. The second framework uses majority voting technique ensemble which combines Light Gradient Boost Machine (LGBM), SVM, and Logistic Regression (LR). This method was 87.9 per cent correct. The methods presented are contrasted with other state-of-the-art solutions and the second solution is found to be the better.

This paper is organized as follows; the second section discusses the related work that used PIDD dataset. Third section explains in details the data mining techniques used in the proposed system. Fourth section shows the PIDD and discusses its details. The fourth section provides the details of the generic framework for diabetes classification. Fifth section shows the experiments and results using the proposed approaches. Finally, sixth section concludes the paper.

II. LITERATURE REVIEW

In this section, we will review the related work with diabetes diagnosis objective and used the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases (Smith et al.,1988).

The paper (Iyer et al., 2015) aim to find solutions to diagnose diabetes disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes (NB) algorithms. The data is preprocessed where missing values are replaced. The data was divided into training set and test set by the cross-validation technique and percentage split technique. 10-fold cross validation is used to prepare training and test data. They used 70% training and 30% testing. The accuracy of decision tree algorithm was 74.9%, 77% for cross-validation technique and percentage split technique respectively. The accuracy of Naïve Bayes classifier was 79.6% for percentage split technique. The authors analyze the bad results by the small data and they would require more data to achieve higher results.

The paper (Choubey et al.,2017) first used NB for the classification on all the attributes and then Genetic Algorithm (GA) used as an attribute selection and NB used on that selected attribute for classification. GA has been used as an attribute (feature) selection method by which four attributes have been selected from eight attributes. They used 70% training and 30% testing. The accuracy of using NB only was 77%. The accuracy of using NB after attribute selection using GA was 78.7%.

The paper (Haritha et al.,2018) first preprocessed the data by eliminating and replacing missing data. Then attribute selection is done using firefly and cuckoo search algorithms. Then the selected attributes are used for classification using K Nearest Neighbor (KNN) and Fuzzy KNN algorithms. Several experiments are done with different training- testing percentage. The accuracy was in the range 63.7% to 77.2%. The highest accuracy was achieved when using cuckoo search for attribute selection and fuzzy KNN for classification with 60% training and 40% testing.

In paper (Barale & Shirke.,2016), the dataset is first preprocessed. They deleted all cases that have two or more of its data missing. Cases with one missing value imputed using method of data imputation KNN. Then boxplot is done to detect outliers which are eliminated from data set leaving it with 498 cases for the modelling

out of 768 cases in the original dataset. K-means clustering algorithm is applied to 498 sample cases for extracting hidden patterns. The misclassified sample cases are removed from data to get final 349 samples. Correctly categorized observations provided as input to the LR, ANN, SVM, and Decision Tree (DT). The dataset is randomly partitioned in two sets with 70% training and 30% testing and 10-fold cross validation also done. The accuracy was in the range 95.7% to 99.3%. The highest accuracy was achieved with LR.

In paper (Deepika & Poonkuzhali.,2015), the dataset is first preprocessed where all cases that have zero value in any of 6 attributes of the dataset is removed. These attributes are Pregnant, Plasma Glucose, Diastolic Blood Pressure, Body Mass Index (BMI). They tried several classification algorithms such as Random Forest, KNN, SVM, DT. They achieved accuracy in the range 77.2% to 90.6% for DT. Then they used an ensemble of KNN and DT which achieved the maximum accuracy 100%.

The paper (Bashir et al.,2016) proposes a model called Hierarchical Majority Voting (HMV) where feature reduction is done using Principal Component Analysis (PCA), then feature extraction is done using Particle Swarm Optimization (PSO), then feature selection using forward selection, backward elimination. The data is then imputed using KNN. Then K-means is used to remove noise. Then a hierarchical majority voting is used for ensemble of classifiers including LR, NB, KNN, SVM and DT. The model is tested over Pima dataset (Smith et al.,1988) and achieved 77.1% accuracy.

The paper (Ateeq & Ganapathy.,2017) proposes a novel framework called Modified Particle Swarm Optimization – Neural Network (MPSO-NN). In that model, the data set is initially pre-processed using the Genetic-Relative Reduct Algorithm. The reduced data is then subject to classification. The two types of ANN, Radial Basis Function Network (RBFN) and Multi-Layer Perceptron Network (MLPN) are compared with each other to figure out which network is best apt network for the classification. Then PSO algorithm is modified as Modified Particle Swarm Optimization (MPSO) Algorithm. The accuracy of applying the framework on Pima dataset (Smith et al.,1988) was 68.8%.

The paper (Gill & Mittal.,2016). proposes Hybrid Prediction Model (HPM). First, the data set is reduced by performing the vertical partitioning, by applying filtration method. SVM and Neural network are combined for making a hybrid model used for classification that gains higher accuracy rate. 230 observations are holdout as testing set. The proposed model achieved 96.1% recognition rate.

The paper (Naik et al.,2016) proposes a model called (SAHS-FLANN) that is self-adaptive harmony search (SAHS) along with gradient descent learning is used with functional link artificial neural network (FLANN) for the task of classification in data mining. The model is used on Pima dataset (Smith et al.,1988) and used 5-fold cross validation. It achieved 80.7% accuracy.

The paper (Akyol & Şen.,2018) first normalized the dataset using min-max normalization. Then they used the Recursive Feature Elimination (RFE), Stability Selection (SS) and Iterative Relief (IR) for feature selection. The data then is used for classification and AdaBoost, Gradient Boosted Trees (GBT) and RF are used for classification. Several experiments have been done, they used different training testing splits 60-40, 70-30 and 80-20 and calculated the average accuracy. The results were in range 70.9% to 73.9%. The highest accuracy is achieved when using SS for feature selection and using AdaBoost for classification.

The paper (Bashir et al.,2014)used ensemble of DT algorithms that's DT based on Information Gain (ID3), DT based on Gain Ratio (C4.5) and DT based on Gini Index (CART) for diagnosis of diabetes from Pima dataset (Smith et al.,1988).The ensembles are Adaboost, Majority Voting, Bagging, Stacking, and Bayesian Boosting techniques. The cross-validation is applied with 10-fold. The accuracy was in the range 68.2% and 74.5%. The highest accuracy is achieved when using Bagging as ensemble technique.

The research (Sundaram.,2018) used Elman Neural network. The Data are divided as 60% for training, 20% for validation and 20% for testing. The resulting classification accuracy is obtained as 95.7%.

The research in (Zhang et al.,2018) used ANN after normalizing the data. The Data are divided as 70% for training, 15% for validation and 15% for testing. The architecture achieved 81.9% accuracy.

The research (Srivastava et al.,2019) used ANN for diabetes diagnosis. The dataset is preprocessed by replacing missing value with the mean. Training and test data divided into a certain number - 688 for training and 80 for testing data. They achieved 92% accuracy.

(Swaroop et al.,2019) proposed a framework that has two stages. The first stage consists of simple Radial Basis Neural Network (RBFN) and simple Probabilistic Neural Network (PNN). The results from both the neural networks are combined and serve as inputs to the second stage classifier which is SVM. The data is partitioned to training that consists of 500 records and testing that consists of 268 records. The proposed model achieved 85.7% accuracy.

III. DATA MINING TECHNIQUES

In this section we review the data mining techniques that is used in the generic framework which are ANN, SVM, Light Gradient Boost Machine (LGBM), and LR.

ANN is a computational model that is based on neural biological networks. Since ANN is conceived on the basis of human biological systems, it learns by an example. Learning is defined as adapting to the synaptic connections between the neurons. This contains an integrated community of artificial neurons as layers and integrates knowledge utilizing a statistical linking strategy. An ANN is an adaptive system based on external or internal information that flows through the network, which changes its structure during the learning phase. The function Activation is used to turn input to output, for example, the function Sigmoid. A cost function for calculating the optimum values of the parameters is used. To find optimal parameters, the Gradient Descent Algorithm or Adam Optimizer is used. The network is implemented several times to boost the model (Gupta& Sedamkar,.2020).

Gradient Boost Machine (GBM) is a gradient boosting framework that uses tree-based learning algorithms that sequentially add predictors to each of its predecessors in order to minimize loss function using gradient descent. It modifies a slow learner sequentially to allow effective use of multiple cores on CPU. LGBM is a GBM variant designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency, lower memory usage, better accuracy, parallel and GPU learning support and the ability to handle large data (Gupta& Sedamkar,.2020).

LR is a tool by which data are categorized into dichotomous tests. It measures the likelihood of an incident occurrence using logistic function which is the probability of occurrence divided by non-occurrence chance. Estimation is done through Maximum Likelihood where model coefficients provide information on the importance of features of the input (Yang & Gong,.2019).

The main building blocks of SVM's are structural risk minimization, nonlinear optimization and spaces induced by duality and kernel features, highlighting the technique with an exact mathematical framework. Several extensions were added to the basic SVM, e.g. for multi-class classification as well as regression and clustering concerns, rendering the methodology commonly relevant in the data mining area (Gill & Mittal.,2016).

IV. PIMA INDIANS DIABETES DATABASE

In this section, we cover in details the PIDD (Smith et al.,1988) from National Institute of Diabetes, Digestive and Kidney Diseases. The age of all of them is at least 21 years old, it has 768 female patient records with 8 features each and a diagnostic attribute. All the functions are numeric. The following Attributes are listed:

- Blood pressure: blood pressure diastolic (mm Hg),
- Skin thickness: Skin thickness of triceps fold (mm)
- Insulin: 2-hour plasma (mu U / ml) hormone
- BMI: index of body mass (weight by kg/ (height by m) ^2)
- Pedigree function in diabetes
- Age: Years-old
- Pregnancies: The number of pregnancies
- Glucose: Concentration of plasma glucose 2 hours in an oral glucose tolerance test

For Normal people and diabetic patients, the class variable is denoted as 0 and 1 respectively. There were 500 samples for normal people, and 268 for patients with diabetes. Figure 1 shows the distribution of normal and diabetes patients.

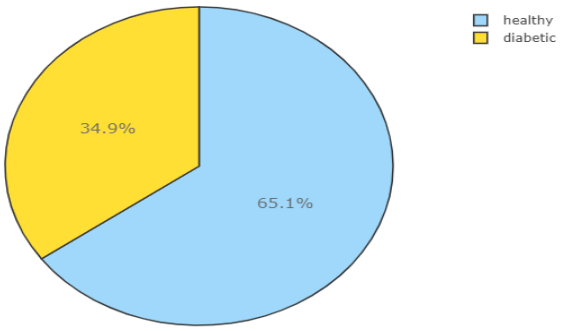

Figure 1: Normal and Diabetic Distribution

V. Proposed Generic Framework

In this section, we would discuss in details the proposed framework for diabetes diagnosis using PIDD dataset (Smith et al.,1988), the proposed framework is clarified in figure 2 which shows the training step and figure 3 which shows the testing step. This section is divided into two main parts. The first part discusses the preprocessing step and the second part discusses the diagnosis step.

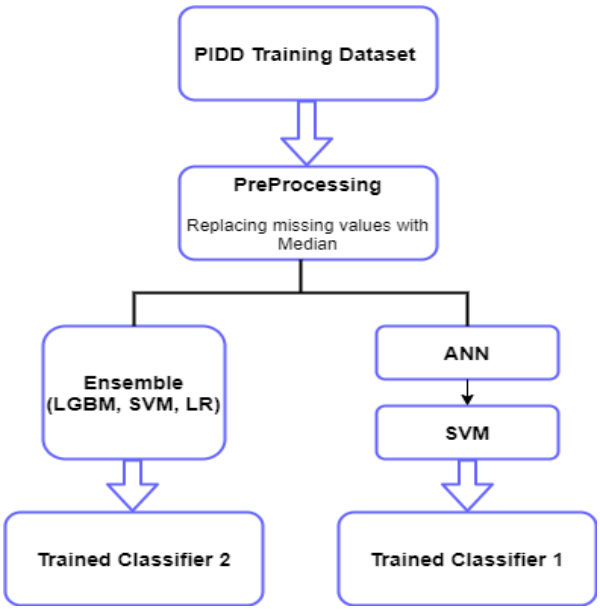

Figure 2: Proposed System (Training)

The PIDD dataset split into training and testing partitions. Figure 2 shows that the training partition is first preprocessed and then is passed to one of two models. Either an ensemble of LGBM, SVM and LR to produce the trained classifier 2, or it is passed to ANN and the output is then passed to SVM to produce the trained classifier 1. Figure 3 shows that the testing partition is first preprocessed and then passed to any of the trained classifiers produced in the training step and the classifier would output the diagnosis result.

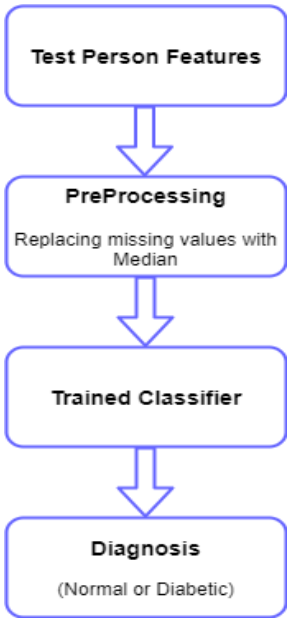

Figure 3: Proposed System (Testing)

The dataset is first preprocessed, the dataset has missing values in five features, and we list them as follows alongside with the percentage of records with missing values and the count of them:

- Insulin: it has 374 records with missing values which makes 48.7%
- Skin Thickness: it has 227 records with missing values which makes 29.56%
- Blood Pressure: it has 35 records with missing values which makes 4.56%
- BMI: it has 11records with missing values which makes 1.43%
- Glucose: it has 5 records with missing values which makes 0.65%

These missing values is replaced with the median value of the same column. Replacing missing value with median is a common technique in dealing with them for preprocessing goal.

The dataset is then classified using two different frameworks that we would discuss in details, namely, hybrid ANN and SVM, and ensemble of LGBM, SVM and LR.

Using the first technique, hybrid ANN and SVM, the dataset is divided into 80% training set and 20% testing set. We define the model starting with the first layer in ANN which is always an input layer where you can define the shape of the initial entry (here are 8 initials from features). Next is the first hidden layer, which depends on the main input layer. Then the second hidden layer. We use several activation functions in the first and second hidden layers that would be discussed in details in the following experiments section. Then we added the output layer. We used the sigmoid as the activation function for the output layer, the main motive why we applied the sigmoid function is that it endures between 0 and 1. Then we compile the model using Adam as an optimizer since it is an adaptive learning rate optimization algorithm specifically designed to train deep neural networks.

Algorithms harness the power of adaptive learning rate methods to find individual learning rates for each parameter (Kingma& Ba,.2014). After completing 100 epochs, with a batch size of 10, the output is fed as input feature for SVM classifier to produce the final output. We use the linear kernel for SVM as the task is a binary classification task where we predict whether the person is normal or diabetic.

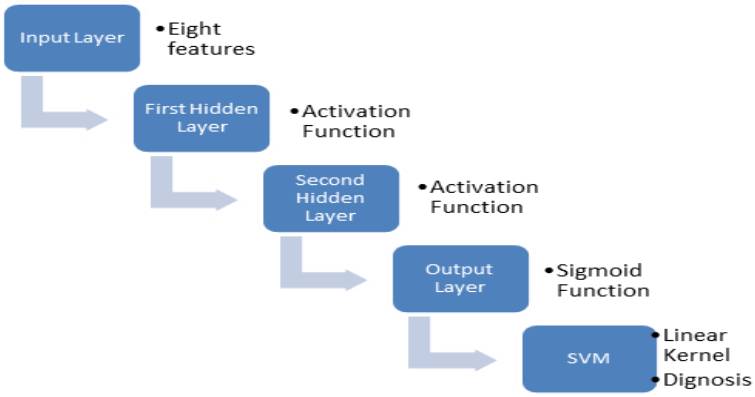

Figure 4: Hybrid ANN and SVM accuracy with different activation functions

The second technique used is the LGBM, SVM, and LR Ensemble. We'll be using random search to find the best hyper parameters for LGBM. Random search is a technique where random hyper parameter combinations are used to find the best solution for the built model. Random Search is generally quicker and more accurate than grid search which calculates all possible combinations. With random grid we decide how many iterations we like. For that ensemble, we use majority soft voting, where voting with the higher score is used as the model's prediction output. With grid search, we look for the best parameters to optimize the voting classifier's accuracy.

VI. Experiments and Results

In the first technique, hybrid ANN and SVM, the dataset is divided into 80% training set and 20% testing set. We have made five experiments on this approach. The main difference between them is the used activation function in the first and second hidden layer. The other parameters of the technique still constant. Here is the list of the used activation functions in the five experiments:

- Linear: it takes inputs, multiplies the weights for each neuron and produces an input-proportional output signal.
- Softmax: standardizes the outputs for each class from 0 to 1, and divides by their total, giving the likelihood that the input value is in a similar class.
- Softplus: The function softplus $f(x) = \ln(1+e^x)$. It is differentiable, and it is able to show its derivative.

- Softsign: The function sofsign (x)=x / abs(x)+1. This generates [-1, +1] scale outputs, which converges polynomially.
- Tanh: Tangent hyperbolic function: $\tanh(x) = (e^{x} – e^{-x}) / (e^{x} + e^{-x})$. This produces [-1, +1] scale outputs, which converges exponentially.

Table 1 shows the activation function used in each experiment alongside with the accuracy achieved with it. Figure 5 clarifies the data in the table. The accuracy is calculated based on the number of true predicted labels divided by the total number of predicted labels. The best accuracy achieved out of the five experiment is the one with tanh activation function with 81.8 accuracy. This is because it converges polynomial which is suitable for our classification task where we classify only two classes. The second-best accuracy is achieved with linear activation function as it is designed for two classes which is the case of our problem. Other activation function accuracies are not good as it is not suitable for our case because they are nonlinear.

Table 1: Hybrid ANN and SVM accuracy with different activation functions

Activation Function	**Accuracy**
Linear	81.2 %
Softmax	75.3 %
Softplus	76.7 %
Softsign	78.6 %
Tanh	**81.8 %**

Figure 5 shows a comparison between the used activation function to produce output in the first and second hidden layer of the proposed ANN. The highest performance of the five studies obtained is the one with 81.8 precision tanh activation feature. This is because it converges polynomially which is suitable for our classification task in which only two classes are classified. With linear activation function, the second-best accuracy is achieved, as it is designed for two classes which is the case of our problem. Other accuracies of the activation function are not good because they are not linear and not suitable for our case.

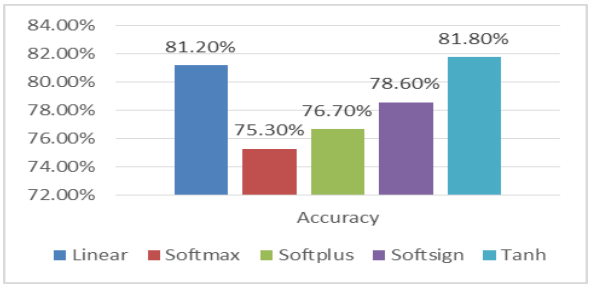

Figure 5: Hybrid ANN and SVM accuracy with different activation functions

In the second technique used, the LGBM, SVM, and LR ensemble. The accuracy of the classifier is calculated using a technique called k fold cross validation. In this technique the dataset is split into k equal partitions and iteratively using one partition as testing set and the reminder as training set and calculating the accuracy. The final accuracy is calculated by taking the average accuracy of all of the iterations. In our approach, we have used 10-fold cross validation. Table 2 shows the accuracy achieved in each iteration of the cross validation. The final average accuracy is 87.9 %. This high accuracy is achieved due to soft voting the ensemble technique used and the search technique in LGBM.

Table 2: 10-fold cross validation average accuracies soft voting of LGBM, SVM, and LR ensemble

Fold	**Accuracy**
1	92.2 %
2	85.7 %
3	84.4 %
4	85.7 %
5	81.8 %

6	89.6 %
7	87%
8	92.2 %
9	89.5 %
10	90.8 %
Average	**87.9 %**

It is obvious that the second approach is better than the first in terms of accuracy. To do the comparison, we selected the papers that have not deleted records of the database as in our case and chose a training/testing partitions that approximates ours. The following table shows a comparison between the proposed systems and other selected state of the art system. We denote the hybrid ANN and SVM that used tanh activation function as first proposed system. Also, the ensemble of LGBM, SVM and LR is denoted as second proposed system. The comparison shows the superiority of the second proposed system. Figure 6 clarifies the data in table 3.

Table 3: Comparison between the proposed system and other system

Approach	**Accuracy**
HMV (Bashir et al.,2016)	77.1 %
MPSO-NN (Ateeq & Ganapathy.,2017)	68.8 %
SAHS-FLANN (Naik et al.,2016)	80.7 %
Ensemble with Bagging (Bashir et al.,2014)	74.5 %
Hybrid RBFN, PNN and SVM (Swaroop et al.,2019)	85.7 %
First Proposed System	81.8 %
Second Proposed System	**87.9 %**

Figure 6 shows a comparison between the two proposed approaches and existing approaches discussed previously in details in the literature reviews sections [(Bashir et al.,2016), (Ateeq & Ganapathy.,2017), (Naik et al.,2016), (Bashir et al.,2014), (Swaroop et al.,2019)]. The first proposed approach is the third best approach. The research (Swaroop et al.,2019) have higher results than the first proposed system as it combines RBFN and PNN and the result is fed to SVM, while our first proposed system uses only ANN and fed it to SVM. The second proposed approach is the first best model as it uses majority voting and combines multiples classifiers including LGBM that has used random and grid search to find the best parameters which made the accuracy of the proposed system the highest.

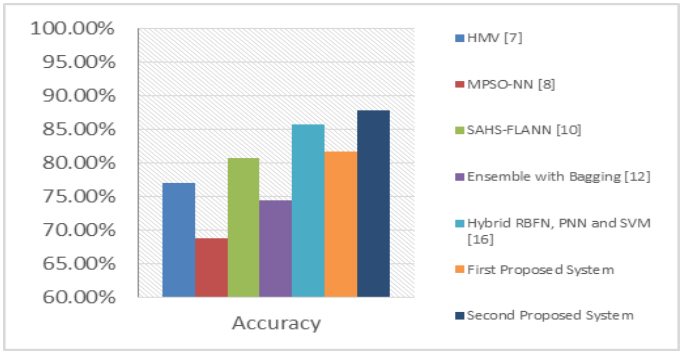

Figure 6: Comparison between the proposed system and other system

VII. Conclusion

The WHO describes diabetes as a multiple etiological metabolic condition defined by persistent hyperglycemia with anomalies in glucose, lipid and protein metabolism triggered by insulin secretion deficiencies, lipid or both. Diabetes is one of the 21st century's most daunting health problems in the world, and affects over 425 million people. Data mining is one of the major techniques which develops and supports medical data research. The aim of this research was to establish a diagnostic framework for diabetes. The PIDD used to function description. Two approaches have been proposed in the study. The first framework uses the

ANN technique and fed production to SVM which yields the diagnostic result. With this approach, five experiments are carried out and the highest accuracy achieved was 81.8%. The second approach employs an ensemble of majority voting techniques which combines LGBM, SVM, and LR. The approach was right at 87.9 per cent. The frameworks presented contrast with other state-of-the-art solutions, and it is found that the second solution is the better.

REFERENCES

[1] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1*.

[2] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

[3] Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2017, February). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)* (pp. 451-455).

[4] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. *International Journal of Applied Engineering Research*, *13*(2), 896-907.

[5] Barale, M. S., & Shirke, D. T. (2016). Cascaded modeling for PIMA Indian diabetes data. *International Journal of Computer Applications*, *139*(11), 1-4.

[6] Deepika, N., & Poonkuzhali, S. (2015). Design of hybrid classifier for prediction of diabetes through feature relevance analysis. *Int. J. Innov. Sci. Eng. Technol*, *2*(10), 788-793.

[7] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: A medical decision support framework using multi-layer classifiers for disease prediction. *Journal of Computational Science*, *13*, 10-25.

[8] Ateeq, K., & Ganapathy, G. (2017). The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO-NN) Algorithm for classifying the Diabetes. *International Journal of Computational Intelligence Research*, *13*(4), 595-614.

[9] Gill, N. S., & Mittal, P. (2016). A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. *J. Theor. Appl. Inf. Technol*, *87*(1), 1-10.

[10] Naik, B., Nayak, J., Behera, H. S., & Abraham, A. (2016). A self adaptive harmony search based functional link higher order ANN for non-linear data classification. *Neurocomputing*, *179*, 69-87.

[11] Akyol, K., & Şen, B. (2018). Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms. *Int. J. Modern Educ. Comput. Sci*, *6*, 10-16.

[12] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In 2014 12th International Conference on Frontiers of Information Technology (pp. 226-231). IEEE.

[13] Sundaram, N. M. (2018). An Improved Elman Neural Network Classifier for classification of Medical Data for Diagnosis of Diabetes. *International Journal of Engineering Science*, *16317*.

[14] Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. (2018). A feed-forward neural network model for the accurate prediction of diabetes mellitus. *International Journal of Scientific and Technology Research*, *7*(8), 151-155.

[15] Srivastava, S., Sharma, L., Sharma, V., Kumar, A., & Darbari, H. (2019). Prediction of Diabetes Using Artificial Neural Network Approach. In *Engineering Vibration, Communication and Information Processing* (pp. 679-687). Springer, Singapore.

[16] Swaroop, K., Cheruku, R., & Edla, D. R. (2019). Cascading of RBFN, PNN and SVM for Improved Type-2 Diabetes Prediction Accuracy. *Australian Journal of Wireless Technologies, Mobility and Security*, *1*(1), 24-27.

[17] Lorenzo, C., Williams, K., Hunt, K. J., & Haffner, S. M. (2007). The National Cholesterol Education Program–Adult Treatment Panel III, International Diabetes Federation, and World Health Organization definitions of the metabolic syndrome as predictors of incident cardiovascular disease and diabetes. *Diabetes care*, *30*(1), 8-13.

[18] Vlassara, H., & Uribarri, J. (2014). Advanced glycation end products (AGE) and diabetes: cause, effect, or both?. *Current diabetes reports*, *14*(1), 453.

[19] Muller, L. M. A. J., Gorter, K. J., Hak, E., Goudzwaard, W. L., Schellevis, F. G., Hoepelman, A. I. M., & Rutten, G. E. H. M. (2005). Increased risk of common infections in patients with type 1 and type 2 diabetes mellitus. *Clinical infectious diseases*, *41*(3), 281-288.

[20] WorldHealthOrganization.http://www.who.int/topics/diabetes_mellitus/en/. Accessed 10 Mar. 2020

[21] Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care*, *27*(5), 1047-1053.

[22] Gupta, S., & Sedamkar, R. R. (2020). Machine Learning for Healthcare: Introduction. In *Machine Learning with Health Care Perspective* (pp. 1-25). Springer, Cham.

[23] Yang, Z., Zhou, Y., & Gong, C. (2019, March). Diagnosis of Diabetes Based on Improved Support Vector Machine and Ensemble Learning. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence* (pp. 177-181).

[24] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.69*.

