A Multilayer Hybrid Machine Learning Model for Diabetes Detection

Sahil Parab[1,*], *Piyush* Rathod[1,**], *Durgesh* Patil[1,***], and *Vishwanath* Chikkareddi[1,****]

[1]Ramrao Adik Institute Of Technology
Navi Mumbai
India

Abstract. Diabetes Detection has been one of the many challenges which is being faced by the medical as well as technological communities. The principles of machine learning and its algorithms is used in order to detect the possibility of a diabetic patient based on their level of glucose concentration , insulin levels and other medically point of view required test reports. The basic diabetes detection model uses Bayesian classification machine learning algorithm, but even though the model is able to detect diabetes, the efficiency is not acceptable at all times because of the drawbacks of the single algorithm of the model. A Hybrid Machine Learning Model is used to overcome the drawbacks produced by a single algorithm model. A Hybrid Model is constructed by implementing multiple applicable machine learning algorithms such as the SVM model and Bayesian's Classification model or any other models in order to overcome drawbacks faced by each other and also provide their mutually contributed efficiency. In a perfect case scenario the new hybrid machine learning model will be able to provide more efficiency as compared to the old Bayesian's classification model.

1 Introduction

Diabetes mellitus also known as diabetes, is a metabolic disorders in which high blood sugar levels are found in the patients body over prolonged period. Frequent urination, increased thirst, and increased hunger can be said to be the symptoms of suspected high sugar level in the patients body. It can lead to many complication cases and issues if the disorder is left untreated.

Diabetes as a metabolic disease have already affected strong number of people worldwide. Its incidence rates can be seen to be increasing exponentially every year. If not treated at the earliest stage of detection, diabetes-related complications and disorders in vital organs of the patients body may turn fatal. Detection of diabetes in the initial phases of the disorder is very crucial for the required treatment routine in order to prevent further complications. From healthcare perspective, machine learning is one of the attractive interest for both medical as well as pharmaceutical researchers.In medical domain where the symptoms of the disorder rely on each other, the machine learning algorithms which are optimized in interpreting the relationships between the features and outcomes is crucial. Heath care needs to understand and use machine learning as the futuristic concept and tool which needs to be deployed in real world using an incremental approach.

Inclination towards the proposed system is due to the lack of accurate and inefficient diabetes detection Systems. This inefficiency in the existing models is due to the complex symptoms that point to different diseases. The crucial part about the idea of the proposed system is that it can detect diabetes based on the probability of a symptom which uses the mathematical model for more accuracy, and the idea sought to achieve the implementation is using multiple machine learning algorithms and summarize their results using a mathematical model.

2 Literature Survey

Significant research has been done so far using the single machine learning algorithm approach in the context of diabetes diagnosis.

The discussion on finding information from medical sources and its importance in order to make an effective medical diagnosis is discussed by the Bayesian Network for Diabetes Prediction[1]. The Naive Bayes model was constructed by applying the classifier on modified preprocessed data set. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3.

The implementation of k-means dependent noise reduction technique[2] is easy to implement, but offers less productivity and needs more computational power which was observed during the study.

While the process that includes data recovery with neural network requires less computation, but offers 85 higher efficiency, the method that can be used to recover data with neural network is more acceptable.

K-means and decision tree[3] have better accuracy than other Type 2 diabetes classification frameworks. In the related studies With the rapidly increasing need for medical data processing, the proposed model can be quite helpful

[*]e-mail: parab.sahil.16ce1061@gmail.com
[**]e-mail: rathod.piyush.16ce2011@gmail.com
[***]e-mail: patil.durgesh.16ce6101@gmail.com
[****]e-mail: vishwanath.c@rait.ac.in

to physicians and doctors for decision making regarding their patients, as they can make more informed judgments by using such an efficient model.

The study carried out by Rubaiat Y.,et al.[4] uses learning algorithms and activation functions which play a crucial role in solving the neural network problem. Compared to ReLU and leaky ReLU ,the activation units show good result for the pima data set as ELU and SELU does not removes the neuronal negative outputs. Lack of adequate accuracy has been seen over some of the algorithms such as multi layer feed forward network[5] for detection of diabetes. So a bit more optimization based on curriculum learning could be applied in the future and check if that improves the outcome. On the basis of distance the data can be rearranged from classifier as rearrangement of data by probabilistic approach doesn't work well.

In Intelligible Support Vector Machines[6] for Diagnosis of Diabetes Mellitus the proposed model uses support vector machines (SVMs) along with additional explanation module for the diagnosis of diabetes mellitus. The model is a new hybrid model for diagnosis in medical, which uses various techniques of machine learning. In particular, for sampling and model building, an unsupervised and supervised learning algorithm is used, respectively, accompanied by a rule-based description section. The model used is SVMs for diabetes diagnosis and prediction, where an additional component for rule-based explanation is used to provide understanding.

The ensemble learning module[7] is used in conjunction with support vector machine[8] that converts black box of SVM decisions into clear and transparent rules. The classification system in ensemble learning utilizes Random Forest (RF) rule induction methodology[8] to establish inexpensive and feasible evaluation criteria for diabetes diagnosis. These rule sets can be considered as a diagnostic second opinion and a tool to screen individuals with undiagnosed diabetes.

M. A. Helal,et al.[9] used several classification algorithm and used them to analyze medical data in order to optimize classifier performance for diabetes and cancer prediction. The data sets used are from machine learning repository at the University of California-Irvine (UCI).The classifiers used are Ensemble Method Decision tree classifier (DTC), K- nearest neighbor (KNN), Naive Bayes(NB), Random Forest (RF), and Perceptron algorithm. For testing purpose K-fold cross validation was used with number of folds k=10. The data set is tested 10 times using K-fold cross which helps classifier to generate higher accuracy. Accuracy is calculated using the values of F-score, recall and precision which are generated using confusion matrix.

S. Lekha,et al.[10] proposes a methodology to improve the overall performance of the applications for real-time detection. This suggests an updated deep learning convolution neural network algorithm combined with vector supporting devices to address the limitations of widely implemented machine learning algorithms. The architecture of the convolution neural network is modified by replacing the multi-layer perceptron classifier fully connected with the SVM algorithm. The dividing technique of the data set into subset, made an optimal classification result. The classification done using several decision tree shows better results, when the data set is divided into subsets using technique such as K-means clustering[11]. Additionally, the classification accuracy of MLNN with LM obtained by the study presented by Temurtas,et al.[12] using correct training was better than those obtained by other studies for the conventional validation method. The multi layer neural network when used with Damped least-squares method yield better result as compared to the other validation method. The modified algorithm optimizes the overall performance and reduces computational complexity compared to the techniques already in use.

3 Working Materials And Algorithms

The data set used in order to test the proposed system is the Pima Indian Diabetes data set of National Institute of Diabetes and Digestive and Kidney Diseases.The data set is composed of total 768 instances and eight numerical attributes represents each of the patient in the data set.
These 8 attributes are:-

- Number of times pregnant
  Min Value : 0 and Max Value : 17
- Plasma glucose concentration of two hours
  Min Value : 0 and Max Value : 199
- Diastolic blood pressure
  Min Value : 0 and Max Value : 122
- Triceps skin fold thickness
  Min Value : 0 and Max Value : 99
- 2-hours Serum insulin
  Min Value : 0 and Max Value : 846
- Body mass index
  Min Value : 0 and Max Value : 67.1
- Diabetes pedigree function
  Min Value : 0.078 and Max Value : 2.42
- Age
  Min Value : 21 and Max Value : 81

The individual machine learning algorithms used as the basic building blocks for the proposed system are:

- **Naive Bayes Classifier**
  Naive Bayes is a technique of classification with a notion that describes all features as separate and unrelated to each other. It determines that the status of a specific characteristic in a class has not been impacted by the status of another characteristic in any way.
  Algorithm:-

```
1.Read the training data set.
2.Standard deviation and mean is calculated of the
  predictor variables in each class.
3.Repeat:
  The Gauss density equation is used to calculate
  probability of fi in each class, until the probabil-
  ity of all predictor variables has been calculated.
4.Calculate the probability for each class.
```

5.Fetch the greatest probability.

Mathematical equation for Naive Bayes Classifier:-

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{1}$$

Here,
P(c | x) is Posterior Probability
P(x | c) is Likelihood Probability
P(c) is Class Prior Probability
P(x) is Predictor Prior Probability

- **Decision Tree Classifier**
  Decision Tree can be classified as the machine learning algorithm for the supervised classification problems.In the proposed hybrid model, the working of the decision tree is to predict required class with respect to the decision rule for the prior data.
  Algorithm:-

  1.Begin the tree with the root node, which contains the complete data set.

  2.Find the best attribute in the data set using Attribute Selection Measure (ASM).

  3.Divide the root node into subsets that contains possible values for the best attributes.

  4.Generate the decision tree node, which contains the best attribute.

  5.Recursively make new decision trees using the subsets of the data set created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Mathematical equation for entropy of a specific node for Decision Tree Classifier:-

$$E(S) = -\sum_{i=1}^{n}(p_i)log(p_i) \tag{2}$$

- **K-Nearest Neighbour clustering**
  KNN can be used for statistical problems both in classification and in regression. It is however more commonly used in industry classification issues.
  Algorithm:-

  1.Data is loaded.

  2.Set the value of k.

  3.Iterate from 1 to total number of training instances, in order to obtain the predicted class.

  4.Euclidean Distance method is used here to calculate distance among the elements of test data and each instances of training data.

  5.In ascending order sort the calculated distances based on distance values.

  6.From the sorted array get top k rows.

  7.Fetch the maximum frequent class of these rows.

  8.Return the predicted class.

The mathemaical equation for the euclidean distance :-

$$D(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{3}$$

- **Support Vector Machine**
  Support Vector Machine comes under the classification-based supervised machine learning model. A SVM's objective is to find the best and the highest-margin separating hyper plane between the given two classes of the training set.
  Algorithm:-

  1.Load the data set.

  2.Define the features and the target.

  3.Before building the SVM algorithm model, Split the data set into test and training set.

  4.Build the support vector machine model by importing the SVC function Sklearn SVM module.

  5.Predict values using the SVM algorithm model.

  6.Evaluate the Support Vector Machine model.

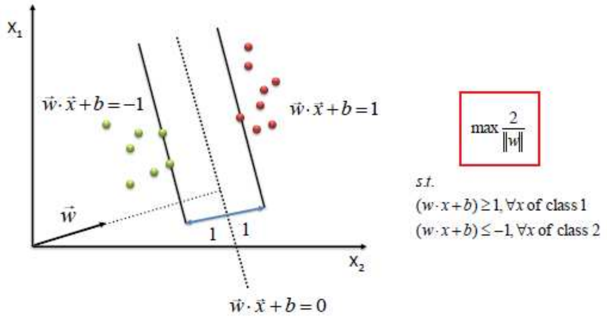

X_1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\max \frac{2}{\|w\|}$
s.t.
$(w\cdot x+b)\geq 1, \forall x$ of class 1
$(w\cdot x+b)\leq -1, \forall x$ of class 2
$\vec{w}$
1
1
X_2
$\vec{w}\cdot\vec{x}+b=0$

Figure 1. Optimal hyper plane to maximize the width of the margin (w)

- **Artificial Neural Networks**
  The Artificial Neural Network (ANN) is known as the method of mathematically representing human neurons by expressing their learning and generalization capabilities.
  Algorithm:-

  1.Load the data set.

  2.To compute the activation values of the Hidden nodes, the inputs from the data set and the linkages are utilized.

  3.In order to compute the activation values of the output nodes, the activation values of hidden nodes found previously and linkages to output are utilized.

  4.Re calibrate all the linkages between output nodes and hidden nodes to find the error rate for the output node.

5. Cascade down the errors to hidden node by using the weights and error calculated at the output node.

6. Weights among input nodes and hidden nodes are re calibrated.

7. Repeat the process till the model meets the criteria for convergence.

8. The activation rate of the output nodes is used to find linkage weights score.

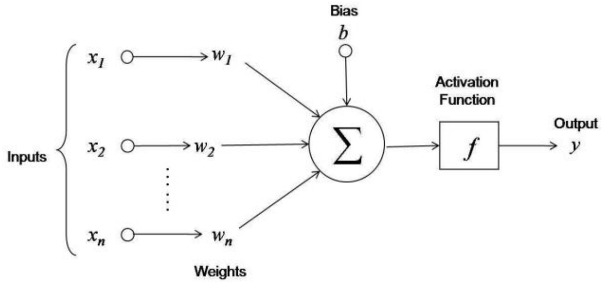


Figure 2. Artificial Neural Network Model

4 Proposed System

The Proposed model is divided into two phases i.e. Training phase and Detection phase.
In training phase, all the machine learning algorithms will be training on the given data set and will be testing their accuracy on the data set.
In detection phase, the machine learning algorithms will be using a fresh input from the data set in order to produce their own classification based on their training from the training phase. These classifications will further be summarized in order to produce final single classification of the detection phase.
It should be observed that the proposed model is not being trained directly on the data set but is indirectly relying on the individual machine learning algorithms and their training results in order to use it in its detection phase.
The basic machine learning algorithms used as the building blocks for the proposed system for our study are :

- Naive Bayes Classifier
- Decision Tree Classifier
- K-Nearest Neighbour Clustering
- Support Vector Machine
- Artificial Neural Networks

These machine learning algorithms are isolated and independent of each other i.e. the algorithms read and process the data without interfering with each other and the results generated by the algorithms does not rely on each other.
The algorithms also does not have any authority to make any modifications to the data set.

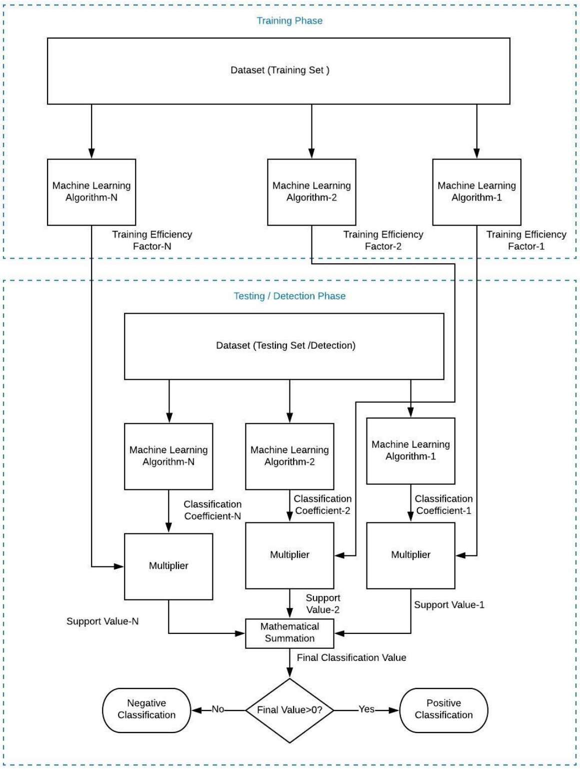

Training Phase
Dataset (Training Set)
Machine Learning Algorithm-N
Machine Learning Algorithm-2
Machine Learning Algorithm-1
Training Efficiency Factor-N
Training Efficiency Factor-2
Training Efficiency Factor-1
Testing / Detection Phase
Dataset (Testing Set /Detection)
Machine Learning Algorithm-N
Machine Learning Algorithm-2
Machine Learning Algorithm-1
Classification Coefficient-N
Classification Coefficient-2
Classification Coefficient-1
Multiplier
Multiplier
Multiplier
Support Value-N
Support Value-2
Support Value-1
Mathematical Summation
Final Classification Value
Negative Classification
No
Final Value>0?
Yes
Positive Classification

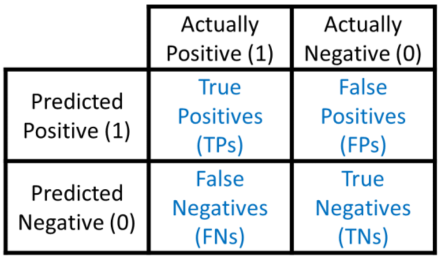
Figure 3. Phase-wise system design for the proposed model

4.1 Preparing Training and Testing data sets

Split the given pima data set into the training set and testing set as per the required distribution.Main focus will be how much training data set is acquired as the maximum training data set will generate a strong training efficiency factor required in further implementation.
The training efficiency factor is defined as the value which represents the efficiency of the algorithm on the training data set.
Consider 75:25 splitting ratio for training and testing.

4.2 Training of algorithm

Individually train every machine learning algorithm in isolation for our hybrid model to obtain the confusion matrix for every algorithm to generate an appropriate training efficiency factor.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 4. Confusion Matrix for machine learning algorithm

As mentioned above, the amount of training data set used influences the training efficiency factor. For every different split ratio of training and testing sets, the training efficiency factor generated is unique.

$$TEF = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

4.3 Detection by individual algorithm

In Detection/Testing Phase, every machine learning algorithm individually operates on the given input and generates a classification coefficient which is binary in nature i.e. it is either generated as +1 or -1. The classification coefficient is +1 if the classification is positive in nature or -1 if it's negative in nature.

$$CC = \begin{cases} +1 & Positive - classification \\ -1 & Negative - classification \end{cases} \tag{5}$$

For every classification coefficient generated by the machine learning algorithm, it is multiplied with corresponding training efficiency factor from training phase. The new value generated is defined as the Support Value for that algorithm for supporting whether the final classification is negative or positive classification.

$$SV = TEF \times CC \tag{6}$$

If the Support value of a algorithm is positive then it is interpreted that the algorithm is supporting the final classification as the positive classification and if it is a negative support value then it supports the final classification as negative classification. It is observed that for every algorithm there is unique support value and the efficiency of every individual algorithm in training phase will make the support value for the final classification of that algorithm either strong or weak towards its nature.

4.4 Summarization of all algorithms

A mathematical model is implemented in order to summarize the support of all the machine learning algorithms for the final classification.

$$FCV = \sum_{i=1}^{n} SV_i \tag{7}$$

As the Hybrid system idea is new and is on basic research stage it will use a summation mathematical model shown in equation (7) where all the support values of the algorithms are summed up with their nature to get a single value which is defined as the final classification value of the hybrid system for the final classification.

The final classification value has two properties: Nature and Magnitude.

The nature of the final classification value will decide the nature of final classification i.e. if the nature is negative then the final classification is negative in nature else its positive in nature.

$$Final - Classification \begin{cases} Positive & FCV > 0 \\ Negative & FCV < 0 \end{cases} \tag{8}$$

The magnitude of final classification value will give how strongly our hybrid system is confident of its classification i.e. if the magnitude is less then the system is barely able to classify and if the magnitude is greater then the system is strongly able to classify.

5 Results

In this research study, our working material was diabetes data set with 768 instances which we first of all divided into two further sets of 700 instances for training and testing of individual algorithms and remaining 68 instances to be used as the live input for the proposed model.

In Training phase of the individual algorithms, the first set of 700 instances are used to obtain TEF for the machine learning algorithms and these are shown in Table-1.

Table 1. Training Efficiency Factors for machine learning algorithms

Algorithm	**TEF**
Naive Bayes Classifier	0.72857
Decision Tree Classifier	0.84762
K-Nearest Neighbour	0.79048
Support Vector Machine	0.77714
Artificial Neural Network	0.90667

In the Detection phase, the remaining 68 instances used as the live input produced the confusion matrix for every individual algorithms and for the proposed system which are presented in the Table-2.

Table 2. Confusion matrix values for individual models and proposed model

Model	**True Positive**	**True Negative**	**False Positive**	**False Negative**
Naive Bayes Classifier	17	34	7	10
Decision Tree Classifier	26	1	40	1
K Nearest Neighbour	27	0	41	0
Support Vector Machine	12	13	28	15
Artificial Neural~ Network	20	16	25	7
Proposed Model	**27**	**6**	**35**	**0**

The above confusion matrix is further used to generate the following result metrics for all the individual models and the proposed model :

- **Accuracy:**
  It is the most common measure of performance for the machine learning model described as the ratio of correctly predicted observations to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

- **Precision:**
  It is described as the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

- **Recall:**
  It is described as the ratio of correctly predicted positive observations to all observations in an actual class.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

The result metrics for all the individual models and the proposed model is presented in Table-3.

Table 3. Result metrics for individual models and proposed model in detection phase

Model	**Result Metric**		
	Accuracy	Precision	Recall
NB Classifier	0.750	0.708	0.629
DT Classifier	0.397	0.393	0.962
KNN Clustering	0.397	0.397	1
SVM Classifier	0.367	0.300	0.444
ANN Model	0.529	0.444	0.740
Proposed Model	**0.485**	**0.435**	**1**

The most beneficial advantage offered by the proposed system is that it is designed in such a way that it considers the classification generated by all the algorithms and not only the most efficient algorithms while discarding the less efficient algorithms during the final classification. Every algorithm is taken into consideration during the final classification based on their efficiency and accuracy in their training phase and their classification coefficient generated in their detection phase.

6 Future Work and Discussion

The hybrid machine learning model being new is still on the basic stage of research and implementation.
Being on the basic stage, the training efficiency factor for the algorithms is considered as their accuracy values obtained from the confusion matrix during the training phase. In future, once the the research on the hybrid system is in progress the training efficiency factor can be generated using complex methods to get more efficient values.
The mathematical model at the basic stage of the research is a basic summation model which directly sum up the support values of all the algorithms to get the final classification value for final classification. In future, for further stages of research a new complex mathematical model will be implemented to get more accurate final classification value for more accurate final classification.
The new hybrid system for binary classification is still in its basic theoretical stage and more research work needs to be done in order to increase its scope of binary classification and real world implementation.

7 Acknowledgment

We take this opportunity to express our profound gratitude and deep regards to our guide Mr. Chikkareddi Vishwanath S. for his exemplary guidance, monitoring and constant encouragement throughout the completion of this study.We are truly grateful to his efforts to improve our understanding towards various concepts and technical skills required in our research study. The blessing, help and guidance given by his time to time shall carry us a long way in the journey of life on which we are about to embark.
We take this privilege to express our sincere thanks to Dr. Mukesh D. Patil, Principal, RAIT for providing the much necessary facilities.We are also thankful to Dr. Leena Ragha, Head of Department of Computer Engineering, RAIT, Nerul Navi Mumbai for her generous support.
Last but not the least we would also like to thank all those who have directly or indirectly helped us in completion of this thesis.

References

[1] Mukesh kumari ,Anshul arora ,Dr.Rajan Vohra , ”Prediction of Diabetes Using Bayesian Network”, International Journal of Computer Science and Information Technologies ,2014.

[2] S. Wei, X. Zhao, and C. Miao, “A comprehensive exploration to the machine learning techniques for diabetes identification,” 2018 IEEE 4th World Forum on Internet of Things (WFIoT), 2018.

[3] W. Chen, S. Chen, H. Zhang, and T. Wu, “A hybrid prediction model for type 2 diabetes using K-means and decision tree,” 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017.

[4] S. Y. Rubaiat, M. M. Rahman, and M. K. Hasan, “Important Feature Selection Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection,” 2018 International Conference on Innovation in Engineering and Technology (ICIET), 2018.

[5] M. Dutt, V. Nunavath, and M. Goodwin, “A Multilayer Feed Forward Neural Network Approach for Diagnosing Diabetes,” 2018 11th International Conference on Developments in eSystems Engineering (DeSE), 2018.

[6] Barakat, N., Bradley, A. P., Barakat, M. N. H. "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus". IEEE Transactions on Information Technology in Biomedicine, 14(4), 1114–1120 (2010).

[7] Han, L., Luo, S., Yu, J., Pan, L. Chen, S. "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis

of Diabetes ". IEEE Journal of Biomedical and Health Informatics 19, 728–734 (2015).

[8] Nongyao Nai-aruna, Rungruttikarn Moungmaia,"Comparison of Classifiers for the Risk of Diabetes Prediction" 7th International Conference on Advances in Information Technology,2015.

[9] M. A. Helal, A. I. Chowdhury, A. Islam, E. Ahmed, M. S. Mahmud, and S. Hossain, “An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction,” 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.

[10] Lekha, S. Suchetha, M. "A Novel 1-D Convolution Neural Network With SVM Architecture for Real-Time Detection Applications". IEEE Sensors Journal 18, 724–731 (2018).

[11] Mustafa S. Kadhm,Ikhlas Watan Ghindawi,Duaa Enteesha Mhawi"An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018)

[12] Temurtas, H., Yumusak, N. Temurtas, F. "A comparative study on diabetes disease diagnosis using neural networks". Expert Systems with Applications 36, 8610–8615 (2009).