DOI: http://dx.doi.org/10.26483/ijarcs.v16i6.7379

ISSN No. 0976-5697

Volume 16, No. 6, November-December 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# MAPREDUCE BASED BIG DATA FRAMEWORK USING DEEP BELIEF NONLINEAR EXPONENTIAL CLASSIFIER FOR DIABETIC DISEASE PREDICTION

S. kamini Pon Seka
Research Scholar,
PG & Research Department of Computer Science,
Government Arts College
(Affiliated to Bharathidasan University, Tiruchirappalli)
Trichy-620022, India

S. Shakila
Head & Associate professor,
PG & Research Department of Computer Science,
Government Arts College
(Affiliated to Bharathidasan University, Tiruchirappalli)
Trichy-620022, India

***Abstract:*** The Healthcare domain is a very distinguished research area with swift technological evolution and surging data progressively. With the intent of extensive healthcare data Big Data Analytics is turning up to be an emerging viewpoint in Healthcare domain. Millions of patients look for treatments globally with numerous procedures. Deep Learning (DL) is an encouraging mechanism that aids in early disease diagnosis and could be beneficial for the practitioners in decision making. This paper aims at building a Deep Learning and MapReduce based Big Data method called, Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) for diabetes prediction. In order to predict it more accurately, this paper proposes a diabetic disease prediction model that combines MapReduce pre-processing, correlated feature selection and classification. Firstly, the diabetic prediction dataset is pre-processed using Batch Normalized Covariate Transpose Propagated MapReduce. Then, combined with two factor correlation analysis between features using correlation coefficient function based on Non-linear Auto Encoding is performed with the optimal feature subset as the feature input. Finally, the Normalized Exponential Classification is used to make robust differentiation between diabetic and non-diabetic via cross entropy as loss function. To evaluate the NACE-NEC methods performance, five different performance metrics, disease prediction time, misclassification rate, precision, recall and accuracy rare validated and analyzed. The NACE-NEC achieved higher performance compared to other state-of-the-art methods on our collected diabetic prediction dataset demonstrating the efficiency of the method in reducing misclassification rate by 40% while improving overall accuracy by 22% and precision extensively.

***Keywords:*** Deep Learning, MapReduce, Big Data, Batch Normalized Covariate Transpose, Non-linear Auto Correlated Encoding

## 1. INTRODUCTION

The swift evolution and widening of big data analytic technology in the healthcare are validating advantageous in disease risk prediction.

Serial Cascaded Convolutional Ensemble Network (SCCEN) was proposed in [1] with the intent of designing efficient diabetes detection systems for the early detection. Initially the required data were gathered from online data source and then fed to optimal feature selection model. Here, fitness based billiards inspired optimization technique was used to select the features optimally. Moreover, the obtained optimal weighted feature was passed to Serial Cascaded Convolutional Ensemble Net work (SCCEN) for early detection with improved accuracy. To address issues related to imbalance factor an attention-enhanced Deep Belief Network (DBN) along with Generative Adversarial Networks (GANs) called, DBN-GAN was proposed in [2] for early diabetes risk prediction. This in turn aided in ensuring models robustness in ascertaining underrepresented cases. In addition hybrid loss function integrating cross-entropy and focal loss was used to boost classification for hard-to-detect instances with improved precision and recall.

A new Recurrent Convolutional Neural Network (RCNN)-based disease risk assessment method was proposed in [3] by employing both structured and unstructured text data from the hospital. Each neuron within convolutional layer received feed forward and obtained recurrent inputs from the preceding unit the nearby unit respectively. Moreover, to ensure fine-grain feature extraction, a step by step recurrent operation was performed on the convoluted output. Finally, data parallelism was also employed during training and testing of proposed model, therefore contributing to enhanced prediction accuracy. Though accuracy was said to be improved, by compromising parameter optimization, a significant amount of training time was found to be involved during risk assessment.

In [4], a deep neural network method was proposed with the objective of predicting blood glucose levels for diabetics in an intermittent level. Here, recurrent neural networks were utilized in an end-to-end fashion along with the patient glucose for significant prediction. As a result, no more feature engineering was necessitated and hence was found to be computationally inexpensive. Though the method was found to be computationally inexpensive, the true positive rate involved in predicting blood glucose was not focused. Prompt and straightforward diabetes diagnosis of individuals employing, hybrid deep learning consisting of genetic algorithm, stacked autoencoder, and Softmax classifier was designed in [5] with higher accuracy.

To prove its efficiency in differentiating between diabetic and non-diabetic cases, logistic regression was applied in [6] that in turn by harnessing key features offered a promising tool. Nevertheless, pre-diabetes prediction remains a demanding issue owing to its biased accuracy and a dearth of explainability. To address on these aspects, a novel hybrid method, combining hyper network with Local Interpretable Model-Agnostic Explanations (LIME)

framework was presented in [7]. This in turn facilitated early intervention.

The growing number of diabetes individuals globally has panicked the medical sector to look for different mechanisms to enhance their medical technologies. Machine learning (ML) and deep learning (DL) algorithms has become active research in designing intelligent and effective diabetes detection systems. A ML based prediction method for diabetes consistently over time was presented in [8] therefore improving accuracy. Due to the noisy and inconsistent nature of data, it can make both laborious and cumbersome in training accurate models. To fill the research gap, a new framework for data modeling based on correlation measures was employed in [9] to process data in an efficient manner for predicting diabetes. An in-depth investigation and influence of the latest ML and DL approaches for diabetes prediction was designed in [10].

Nowadays, diabetes is one of the most prevalent and chronic diseases owing to certain amount of complications globally. The early diabetes detection is very paramount for its timely therapy due to the reason that it can be stopped during disease progression. In [11], enhanced deep neural network was applied to not only predict the diabetes occurrence but also to ascertain disease type with improved training accuracy. Yet another deep neural network with multi layer perceptron was presented in [12], therefore boosting precision and accuracy. A comparative analysis of DL and statistical methods were designed in [13] to focus on computational cost and accuracy rate. A systematic review on the application of ML and DL techniques for diabetes detection was investigated in [14]. In [15], comparison of ML algorithms for early diabetes prediction was designed focusing on the accuracy and precision aspects.

### 1.1 Problem statement

The problem for differentiating between diabetic and non-diabetic involving vast amount of patient data necessitates addressing the issues connected with the early intervention and personalized treatment plans accurately and precisely with minimal misclassification. To handle this issue, there is a requirement for robust and sophisticated diabetic disease prediction methods. The focus is on using deep learning framework tailored with MapReduce to ensure accurate and precise diabetic disease prediction and ensure early treatment.

### 1.2 Research gap

In spite of the elaborate research works on diabetic disease prediction for vast amounts of patient data, along with deep learning applications, gaps in research persist. One area with room for exploration is the use of deep learning and MapReduce to process samples based on linear and non-linear factors while selecting the most relevant features. Moreover, the interpretability of deep learning models with MapReduce in diabetic disease prediction is in the inception stage, with limited methodologies for comprehensively understanding feature significance. Addressing these gaps is necessary for boosting diabetic disease prediction in response to increasing samples.

### 1.3 Contributions of the work

To address on the above said issues, like handling training time and misclassification rate with focus on accurate disease detection results using the Diabetes prediction dataset, a method called, Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) is designed. The contributions of the NACE-NEC are listed as given below.

- To design a robust diabetic disease prediction method, NACE-NEC is designed based on pre-processing, feature extraction and classification via one input, three hidden layers and one output layer.
- To minimize training time involved in the overall process of diabetic disease prediction, Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing is used.
- To design Non-linear Auto Correlated Encoding Feature Selection model that by applying correlation coefficient function selects highly correlated features, blood_glucose_level, Hba1c_level, age, hypertension, heart disease and smoking_history whereas by using non-linear auto encoding selects BMI possessing non-linearity however of most significant for prediction that in turn aids in minimizing the misclassification rate in an extensive manner.
- To improve accuracy by applying Normalized Exponential Classification for diabetic disease prediction that along with the cross entropy as loss function aids in minimizing false positive and false negative samples in a significant manner.
- Finally, complete experimental evaluation is performed with five different performance metrics, training time, misclassification rate, precision, recall and accuracy to illustrate the proposed NACE-NEC method over traditional methods.

### 1.4 Organization of the paper

This paper is organized as follows: In Section 2, we motivate our study by setting in motion the background and related work involving diabetic disease prediction using deep learning techniques. In Section 3, we present the details of our proposed Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) method for efficient diabetic disease prediction. Experimental results along with a comprehensive discussion with the traditional methods using table and graphical representations are provided in Section 4. Also a detailed case study and inferences are included ensuring qualitative analysis. Following which a detailed discussion with the conventional methods using table and graphical representations are provided to ensure quantitative anlaysis. Finally, concluding remark is included in Section 5.

## 2. RELATED WORKS

Given the increasing population, it is requisite to design mechanisms to improve health and circumvent increasing concerns globally. With the advancement seen in scientific research, the evolution of such system is becoming more effective. Hence early diabetes detection is very paramount due to the reason that if missed with timely treatment can stop the progression of the disease. A systematic review on diabetes progression employing ML and DL was investigated in [16]. With the disease now decreased to mid-twenties and given the high prevalence, it is required to address this diabetes problem efficiently. Issues referring to data scarcity and model deployment were discussed in [17]. A method for early diagnosing by

parallelizing polynomial kernel vector with MapReduce was designed in [18] for precise diabetes disease prediction.

Diabetes is a widespread disease globally that considerably reduces the quality of life and can even result in fatalities owing to its difficulties. Stacked auto encoders in DL were applied in [19] for reducing risk factors involved in wrong prediction. Diverse ML algorithm was employed in [20] for early diabetes diagnosis with improved accuracy. Yet another method to focus on accuracy employing dual teacher knowledge distillation and feature enhancement was presented in [21] ensuring improved classification performance. Early diagnosis employing generative adversarial networks and radial basis neural network were applied in [22] with improved test accuracy.

An efficient medical decision system for predicting diabetes employing Deep Neural Network (DNN) was presented in [23]. By employing this DNN not only improved accuracy but also reduced training time considerably. A variety of ML techniques were applied in [24] to enhance diabetes prediction performance. Yet another mechanism for early diagnosis employing stacking model based on genetic algorithm and XGboost techniques were employed in [25] to improve the model's predictive accuracy. A method to ensure faster convergence and higher prediction accuracy employing wrapper based feature selection along with multilayer perceptron was proposed in [26].

An enhanced artificial neural network method training employing backpropagation scaled conjugate gradient neural network was presented in [27] with the intent of predicting diabetes in efficient manner. Nevertheless, prevailing prediction methods struggle to capture accurately indispensable features of nonlinear data. To address on this issue, diabetes prediction method integrating Kendall's correlation coefficient and attention mechanism was presented in [28]. Finally, a deep neural network based on the self attention mechanism was constructed that in turn improved the method's predictive performance. Also temporal features play a major role in disease diagnosis. In [29] artificial intelligence with temporal features was introduced in predicting diabetes with minimal time.

Motivated by the above literature studies, where certain works concentrated on precision and accuracy whereas the other focused on the misclassification rate. To concentrate on these aspects, in this work a robust diabetes disease prediction method called Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) is provided in the following sub-sections.

## 3. MATERIALS AND METHODOLOGY

Big data analytics and deep learning is transforming healthcare by enabling integration of extensive amounts of information to boost patient care, optimize operations and ensuring evolution in research. With the abundance of healthcare dataset and advancements in DL models systems in the recent years are well equipped in diagnosing diabetic disease at an early stage. In this work, Deep Learning and MapReduce based Big Data method called, Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) is designed for robust diabetes prediction. Figure 1 shows the structure of NACE-NEC method.

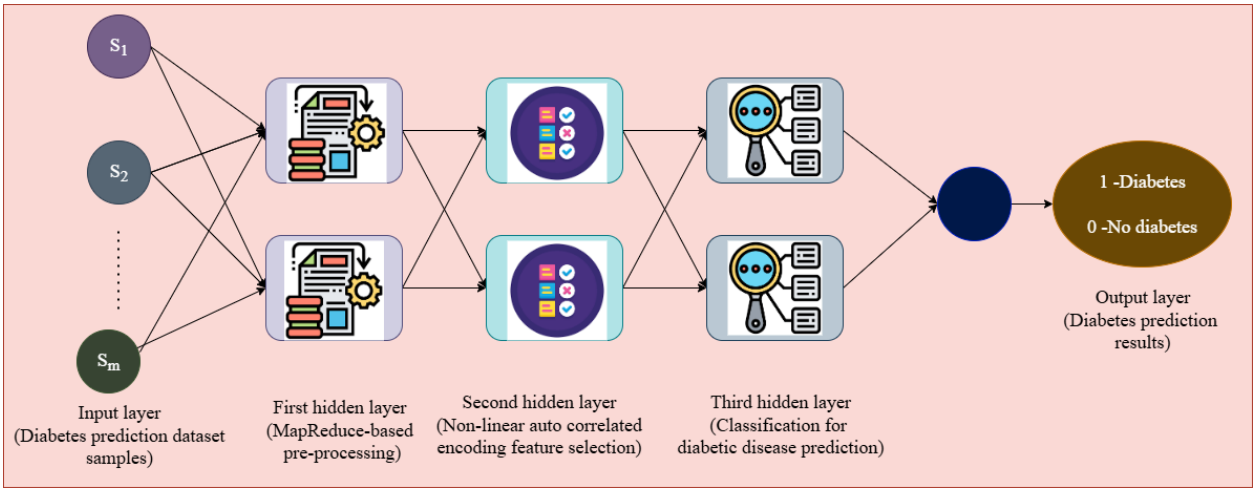


**Figure 1 Structure of NACE-NEC method**

As shown in the above figure, the NACE-NEC method includes three layers. They are input layer, three hidden layers and the output layer. The samples from raw diabetic prediction dataset are provided as input in the input layer. In the first hidden layer, Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing is performed. Here, the overall samples are split into chunks and processed parallel by mapper task. Then, key-value pairs grouped are sent to reducer tasks, therefore generating processed samples. In the second hidden layer, Non-linear Auto Correlated Encoding Feature Selection model is utilized. The purpose of using this model is that it make sure that not only error is back-propagated therefore obtaining the best features to be learned in an efficient manner. Finally, Normalized Exponential Function is employed in the third hidden layer to early disease prediction and normalizes the output, ensuring significant true positive rate.

### 3.1 Data collection

The Diabetes prediction dataset extracted from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset comprises a collection of patient's medical and demographic data in addition to their diabetes status (positive or negative). The ten features in the dataset

are ID, age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level and target variable. With the dataset details provided the MapReduce framework with big data is provided in the following sub-sections.

### 3.2 Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing

MapReduce-based pre-processing influences the MapReduce programming model to carry out data preparation tasks on large datasets in a distributed and parallel manner. This procedure is specifically advantageous for big data environments where conventional single-machine processing is non-viable. In this work, Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing is employed. Figure 2 shows the block diagram of Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing model. The data pre-processing process on MapReduce consists of two main tasks. They are one-hot encoding and normalization.

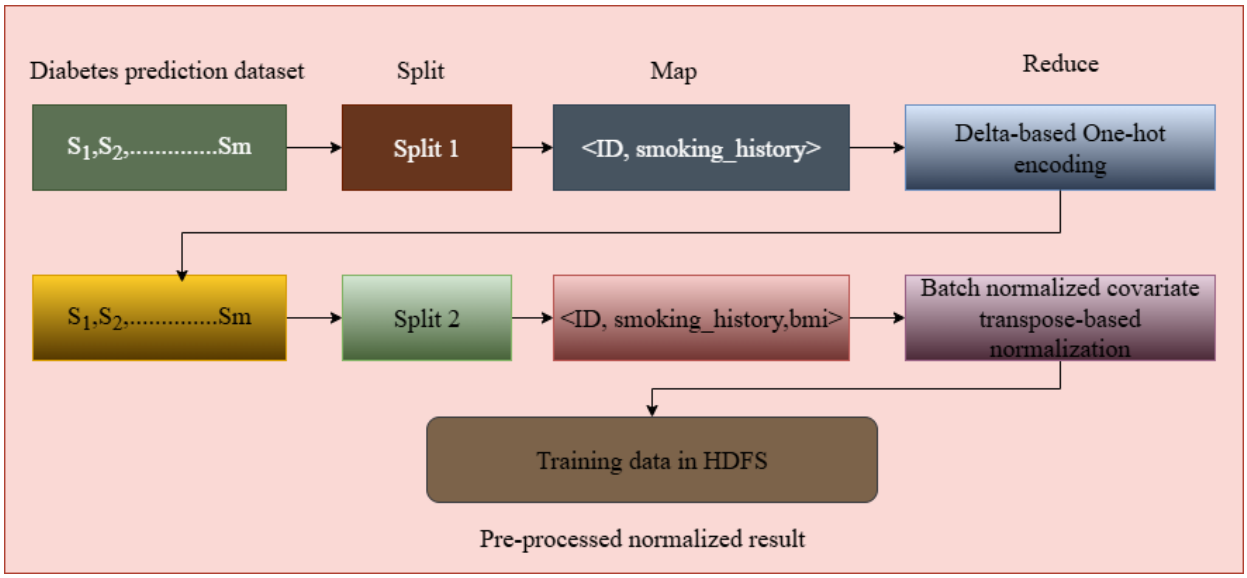


**Figure 2 Block diagram of Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing model**

As shown in the above figure first, the Diabetes Prediction dataset collected from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset is subjected to Delta-based One-hot encoding. Following which the results are normalized using Batch Normalized Covariate Transpose function to obtain pre-processed results for further processing. Here, a key-value pair in the first Map task is defined, where key 1 represents the ID and value 1 the smoking sample values. Input to the Reduce task, in the Reduce stage, by performing one-hot encoding on smoking. In the second Map task, the BMI values are read and the key-value pair is defined, where the one-hot encoded results of smoking is used as key2 and value2 represents the normalized BMI values and write to the resultant value to Hadoop Distributed File System (HDFS)further processing.

First, with the raw data obtained from Diabetes Prediction dataset Delta-based One-hot encoding is applied for converting categorical data into numerical format for easy understanding. The binarization of categorical data into numerical format is achieved using Delta-based One-hot encoding. Let us suppose that we wish to solve integer optimization problem with '$m$' integer variables '$S_i, i = 1,2, \dots, m$' is mathematically represented as given below.

$$Res_i = argmin \sum_{i=1}^{m} J_{i,i+1}\delta(S_i, S_{i+1}) \quad (1)$$

From the above equation (1), '$S_i \in (1,2,..F)$', '$F$' denotes the number of features, '$J_{i,i+1}$' denoting the association between '$S_i$' and '$S_{i+1}$' mapped via Kronecker Delta function '$\delta$'. Then, the integer variables '$\{S_i\}_{i=1,2,\dots m}$'is binarized by one-hot encoding as given below.

$$argmin \sum_{i=1}^{m} J_{i,i+1} \sum_{f=1}^{F} a_i^{(F)} a_{i+1}^{(F)}, such\ that \sum_{f=1}^{F} a_i^{(F)} = 1 \quad (2)$$

$$Temp = \{(S_1, Res_1), (S_2, Res_2), \dots, (S_m, Res_m)\} \quad (3)$$

From the above equation (2) '$a_i^{(F)} \in (0,1)$' denotes binary variable assigned to feature '$f$' of '$S_i$' and '$a_i^{(F)} = 1$' represents the feature '$f$' is selected for '$S_i$'. Finally the one-hot encoded results are stored in '$Temp$' for further processing.

Following which, Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing is applied. Upon comparison to Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing, the conventional batch normalization heavily depends on batch statistics for normalization during training that makes evaluation of mean and standard deviation of samples inaccurate owing to shifting parameter values. Also issue with conventional batch normalization is that it cannot be utilized with batch size of 1 during the training process. So by employing Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing does not depend on batch statistics. This is because by applying MapReduce with Batch Normalization and Covariate Transpose perform data preparation in a distributed and parallel manner. Hence, by evaluating mean and standard deviation in every independent chunks during Map phase and shuffle during Reduce phase based on key, thus being computationally faster. First, a strategy for normalizing the data in such a manner the gradient update step accounts for normalization. This is mathematically represented as given below.

$$Temp_i' = \frac{Temp_i - E_B[Temp_i]}{\sqrt{Var_B(Temp_i)}} \quad (4)$$

From the above equation (4), '$Temp_i$' denotes the '$i-th$' element of '$Temp$' and the expectation '$E_B$' is calculated over training mini-batch '$B$'. Secondly, this normalization is performed for each sample independently employing mini-batch statistics.

$$PD = \alpha Temp_i' + \beta = BN_{\alpha\beta}(Temp_i) \quad (5)$$

From the above results the pre-processed normalized data '$PD$' are obtained for further processing via the learnable parameters '$\alpha = 1$', '$\beta = 0$' respectively. The pseudo code representation of Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing is given below.

**Input**: Dataset '$DS$', Samples '$S = \{S_1, S_2, \ldots, S_m\}$', Features '$F = \{F_1, F_2, \ldots, F_n\}$'
**Output**: Computationally-efficient pre-processing 'pre-processed samples '$PS$''
1: **Initialize** '$m = 100000$', '$n = 9$', '$\alpha = 1$', '$\beta = 0$'
2: **Begin**
**//Delta One-hot Encoding [Smoking_History feature]**
3: **For** each Dataset '$DS$' with Samples '$S$'and Features '$F$'
4: Perform One-hot Encoding according to (1) and (2)
5: **Return** one-hot encoded results '$Temp$'
6: **End for**
**//Batch Normalized Covariate Transpose-based Normalization [BMI feature]**
7: **For** each Dataset '$DS$' with Samples '$S$', Features '$F$' and one-hot encoded results '$Temp$'
8: Perform normalization in such a way that gradient update step accounts for normalization according to (4)
9: Perform normalization independently employing mini-batch statistics according to (5)
10: **Return** pre-processed samples '$PS$'
11: **End for**
12: **End**

**Algorithm 1 Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing**

As given in the above algorithm with the big data revolution in healthcare though too large for traditional systems has paved the way for Map reduce mechanism. With this objective, the above Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing process involves Delta One-hot Encoding for Smoking_History feature and Batch Normalized Covariate Transpose-based Normalization for BMI feature to be performed via Map Reduce mechanism. By applying the above two functions first, the overall samples are split into chunks and processed in parallel by mapper task. Following which with the aid of intermediate key-value pairs generated for both Smoking_History feature and BMI feature in the dataset are shuffled and grouped via ID. Finally, the grouped key-value pairs generated of both Smoking_History feature and BMI feature are sent to reducer tasks. This in turn aids in further minimizing training process.

### 3.3 Non-linear Auto Correlated Encoding Feature Selection

Feature selection models are aimed at minimizing unimportant features and concentrating on the features that contribute to the most predictable feature of the target feature. In the second hidden layer, Non-linear Auto Correlated Encoding Feature Selection model is utilized with the objective of reducing the misclassification error in such a manner that the error are easily back-propagated, therefore selecting the best features to be learned for further processing. The traditional auto encoder refers to a symmetric neural network trained in such a manner so as to copy its input to its output utilizing a hidden layer as shown in figure 3.

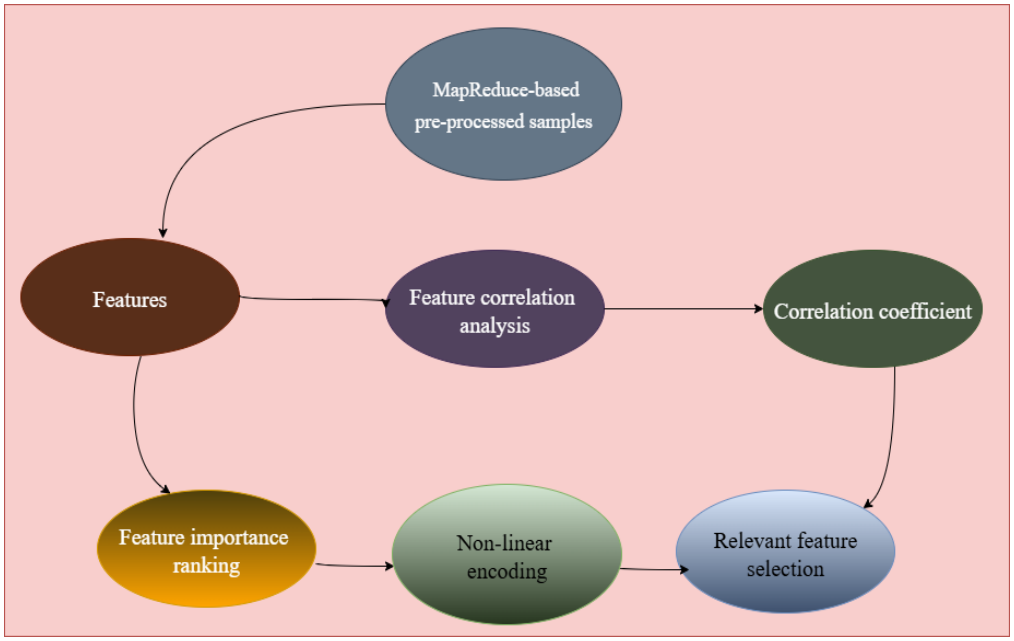


**Figure 3 Block diagram of Non-linear Auto Correlated Encoding Feature Selection**

The methods for selecting features by means of optimization include assessing the association between each feature and the target feature and selecting the input features that have the strongest correlation with the target feature. In this study, Correlation Coefficient function is used to perform the correlation analysis between features and Non-linear Auto Encoding is used to perform feature importance ranking with the intent of extracting the best features from the dataset. To start with the correlation coefficient of random features '$A \in F$' and '$B \in F$' is represented as given below.

$$\rho = \frac{cov(A,B)}{\sqrt{Var(A)Var(B)}} \tag{6}$$

From the above equation (6) '$cov$' represents the covariance of random features '$A \in F$' and '$B \in F$' whereas '$Var$' denotes the variance of random features separately denoted as '$A \in F$' and '$B \in F$' respectively. The evaluation of the correlation coefficient function between the features '$A$' and '$B$' is mathematically represented as given below.

$$r = \frac{\sum_{i=1}^{n} (A_i - A')(B_i - B')}{\sqrt{\sum_{i=1}^{n} (A_i - A')^2 \sum_{i=1}^{n} (B_i - B')^2}} \tag{7}$$

From the above equation (7) '$n$' denotes the total feature size, '$A_i$', '$B_i$' represents the observations of feature '$A$' and '$B$' and finally '$A'$', '$B'$' denoting the mean feature size of '$A$' and '$B$' respectively. The correlation coefficient is a numerical measure with values between '$-1$' and '$1$'. The linear relationship between '$A$' and '$B$'is said to be strong if the correlation coefficient resultant value is close to '$-1$' or '$1$'. On the other hand, correlation of '$0$', mean that there is no linear relationship between '$A$' and '$B$'. In spite of the above correlation coefficient capturing linear associations, by employing non-linear encoding function transform data to uncover hidden non-linear associations that can be ascertained by correlation analysis. This permits in selecting the features that are strongly associated to the target variable, even if the association is not found to be linear. In the proposed method, to uncover this aspect, feature importance ranking is performed using non-linear auto encoding. For data compression the hidden layer is constrained in such a manner that the number of neurons in the hidden layer is smaller upon comparison to the number of neurons in the input layer. This constraint intensifies the auto encoder to select the most salient features in lower dimensionality upon comparison to the input space. Figure 4 shows the block diagram of non-linear auto encoding-based feature importance ranking.

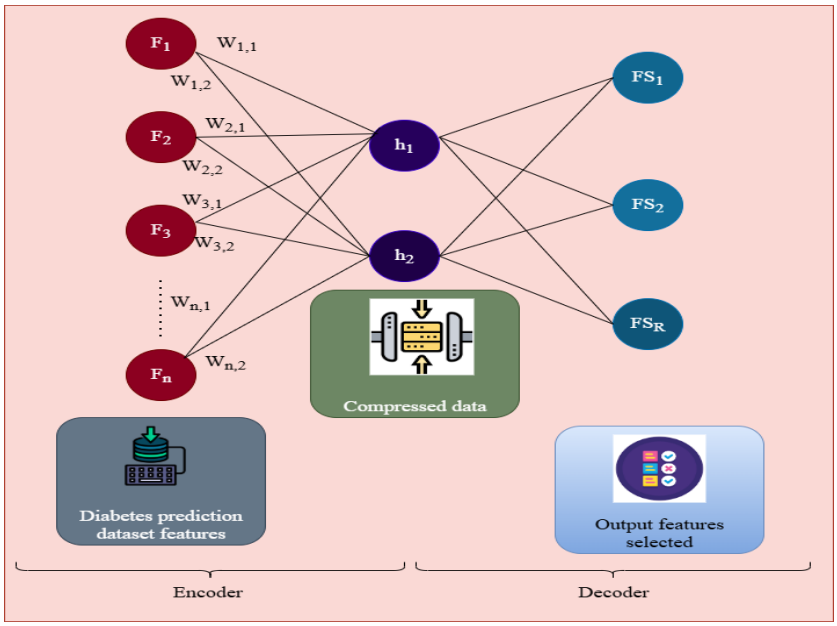


**Figure 4 Block diagram of non-linear auto encoding-based feature importance ranking**

As shown in the above figure, dimensionality reduction has occurred because the '$n - dimensional$' input vector '$(F_1, F_2, \dots, F_n)$' receives a compressed hidden representation '$(h_1, h_2)$' in two dimensions, referred to as the latent space. Moreover all the input features contribute to latent space representations uniformly. For example, in case of '$w_{3,1}, w_{3,2} = 0$', the summarization can be made that the feature '$F_3$' is a redundant feature and hence is eliminated from the input feature set without any negative effects in further processing. Moreover, if the input layer has '$n$' feature nodes, the hidden layer has '$k$' nodes and the accumulated weight of the link between input feature node '$i$' and the hidden layer node '$j$' is '$W(i,j)$' after training, then the weight is measured via forward propagation as given below.

$$W(i) = \sum_{j=1}^{k} \frac{W(i,j)}{k} \tag{8}$$

$$for^{(l)} = \sigma\left(FS^{(l)}\right) \tag{9}$$

With the above results (8) the vector of feature weights '$F = \big(W(1), W(2), \dots W(n)\big)$' is constructed for further processing. This in turn make certain that the Setting this vector into descending arrangement provides an assessed mechanism to weigh up the correlative significance of one feature to another and to select the most important features. On the other hand, in backward propagation, the

evaluated results are compared with the aid of a loss function to measure the error that in then propagated backward to fine-tune the weights in an iterative fashion until error rate reaches an acceptable level.

$$\delta^{(k)} = \frac{\partial L}{\partial FS^{(l)}} = \delta^{(l+1)} \cdot \frac{\partial FS^{(l+1)}}{\partial for^{(l)}} \cdot \frac{\partial for^{(l)}}{\partial FS^{(l)}} \quad (10)$$

By propagating backward to fine-tune the weights in an iterative fashion in turn make sure that the misclassification rate are straight forwardly back-propagated so that the best features are said to be learned in an efficient manner. The pseudo code representation of Non-linear Auto Correlated Encoding Feature Selection is given below.

**Input**: Dataset '$DS$', Features '$F = \{F_1, F_2, \dots, F_n\}$'
**Output**: error minimized features selected '$FS$'
1: **Initialize** '$m = 100000$', '$n = 9$', pre-processed samples '$PS$'
2: **Begin**
3: **For** each Dataset '$DS$' with Features '$F = \{F_1, F_2, \dots, F_n\}$' and pre-processed samples '$PS$'
//**Correlation analysis based on Correlation Coefficient function**
4: Evaluate correlation coefficient of random features according to (6)
5: Evaluate correlation coefficient function between features '$A$' and '$B$' according to (7)
6: **If** '$r\ is\ close\ to -1\ or\ 1$'
7: **Then** features are said to be relevant and selected for further processing
8: **Return** correlation analysis results [blood_gluclose_level, Hba1c_level, age, hypertension, heart_disease, smoking_history]
9: **End if**
10: **If** '$r\ is\ equal\ to\ 0$'
11: **Then** features are not said to be relevant and go to step 15
12: **End if**
//**Feature importance ranking based on Non-linear Auto Encoding**
13: Measure feature importance ranking according to (8), (9) and (10)
14: Return non-linear results [blood_gluclose_level, Hba1c_level, age, hypertension, heart_disease, smoking_history, BMI]
14: **Return** relevant features selected '$FS$'
15: **End for**
16: **End**

**Algorithm 2 Non-linear Auto Correlated Encoding Feature Selection**

When analyzing healthcare data, it is important to determine which feature is significant, so feature selection models are more pertinent for processing such healthcare data. The proposed feature selection procedure as given above consists of two steps. In the first step, feature selection is performed by selecting the most significant feature by analyzing the correlation between features using correlation coefficient function. In the second step, the features obtained via the first step are combined into a single set of significant features by determining feature importance ranking using non-linear auto encoding. This in turn aids in reducing misclassification rate involved in diabetic prediction.

### 3.4 Normalized Exponential Classification for diabetic disease prediction

Finally in this section, with the pre-processed samples and features selected as input, Normalized Exponential Classification is applied for diabetic disease prediction. Then with binary classification problem (i.e. diabetics '1' or not diabetics '0') influencing the Normalized Exponential function as the activation function. This Normalized Exponential function is mathematically formulated as given below.

$$Res_i = \frac{expexp\ (PS_i)}{\sum_{i=1}^{l}\ (PS_j)} \quad (11)$$

From the above equation (11), '$l$' represents the classes (either '1' or '0') with '$Res_i$' denoting the '$i-th$' element of the output vector (i.e. '$Res_i \in [0,1]$') and '$PS_i$' representing the '$i-th$' element of the input vector respectively. In order to acquire the gradient of the Normalized Exponential function cross entropy is used as the loss function. This is defined as given below.

$$L = -\sum_{i=1}^{l} \quad Act_i\ log\ log\ (Pred_i) \quad (12)$$

From the above equation (12), '$l$' denotes the binary classes (i.e. '1' for diabetic and '0' for non-diabetic), '$Act_i$' representing the actual label for '$i-th$' sample and '$Pred_i$' denoting the predicting label for the '$i-th$' sample respectively. The pseudo code representation of Normalized Exponential Classification for diabetic disease prediction is given below.

**Input**: Dataset '$DS$'
**Output**: accurate and precise diabetic disease prediction
1: **Initialize** '$m = 100000$', '$n = 9$', pre-processed samples '$PS$', features selected '$FS$'
2: **Begin**
3: **For** each Dataset '$DS$' with pre-processed samples '$PS$' and features selected '$FS$'
4: Evaluate Normalized Exponential function according to (11)
5: Measure cross entropy as loss function according to (12)

6: **If** '$age \geq 45$' and '$Hba1c - level \geq 6.5$' and '$blood - glucose - level \geq 126$'
7: **Then** '$Res_i = 1$'
8: Sample identified with diabetes
9: **Else**
10: Sample identified with no diabetes
11: **End if**
12: **End for**
13: **End**

**Algorithm 3 Normalized Exponential Classification for diabetic disease prediction**

As given in the above algorithm, first, pre-processed samples and relevant features selected were obtained as input for classification. Here, Normalized Exponential function converts a vector of raw scores into a probability distribution. The advantage of using this function is that it produces non-negative outputs that sum to one, making certain that each classified results represents a valid probability. As a result precise and accurate differentiation between classes is made in a significant manner. Also by using the cross entropy as the loss function while obtaining gradient of Normalized Exponential function the false negative rate was reduced therefore improving the overall recall rate significantly.

## 4. EXPERIMENTAL RESULTS

In this section, performance of proposed Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) method is evaluated b based on several evaluation metrics. Besides, comparison between our proposed method and conventional methods, Serial Cascaded Convolutional Ensemble Net work (SCCEN) [1] and Deep Belief Network with Generative Adversarial Networks (DBN-GAN) [2] has been conducted aiming to signify the superiority of our method. To perform simulation, Diabetes Prediction dataset extracted from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset is employed and fair comparison is made for all the three methods using Java language.

### 4.1 Case study and inferences

In this section the qualitative analysis of proposed Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) method for diabetes prediction is detail with inferences. To perform simulations samples were obtained from Diabetes prediction dataset. With the raw 100000 samples obtained from the dataset, pre-processing was performed by applying Batch Normalized Covariate Transpose Propagated function. Figure 5 shows the results of pre-processing.

Delta-based One-hot encoding MapReduce results

| | ID | gender | age | hypertension | heart_disease | bmi | HbA1c_level | blood_glucose_level | diabetes | smoking_history_never | smoking_history_No Info | smoking_history_current | smoking_history_former | smoking_history_ever | smoking_history_not current |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Female | 80.0 | 0 | 1 | 25.19 | 6.6 | 140 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | Female | 54.0 | 0 | 0 | 27.32 | 6.6 | 80 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 3 | Male | 28.0 | 0 | 0 | 27.32 | 5.7 | 158 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | Female | 36.0 | 0 | 0 | 23.45 | 5.0 | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 4 | 5 | Male | 76.0 | 1 | 1 | 20.14 | 4.8 | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 99995 | 99996 | Female | 80.0 | 0 | 0 | 27.32 | 6.2 | 90 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 99996 | 99997 | Female | 2.0 | 0 | 0 | 17.37 | 6.5 | 100 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 99997 | 99998 | Male | 66.0 | 0 | 0 | 27.83 | 5.7 | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 99998 | 99999 | Female | 24.0 | 0 | 0 | 35.42 | 4.0 | 100 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 99999 | 100000 | Female | 57.0 | 0 | 0 | 22.43 | 6.6 | 90 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

[100000 rows x 15 columns]

Batch Normalized Covariate Transpose MapReduce- Results

| | ID | gender | age | hypertension | heart_disease | bmi | HbA1c_level | blood_glucose_level | diabetes | smoking_history_never | smoking_history_No Info | smoking_history_current | smoking_history_former | smoking_history_ever | smoking_history_not current |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Female | 80.0 | 0 | 1 | 0.177171 | 6.6 | 140 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | Female | 54.0 | 0 | 0 | 0.202031 | 5.7 | 80 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 3 | Male | 28.0 | 0 | 0 | 0.202031 | 5.0 | 158 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | Female | 36.0 | 0 | 0 | 0.156863 | 4.8 | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 4 | 5 | Male | 76.0 | 1 | 1 | 0.118231 | ... | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 99995 | 99996 | Female | 80.0 | 0 | 0 | 0.202031 | 6.2 | 90 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 99996 | 99997 | Female | 2.0 | 0 | 0 | 0.085901 | 6.5 | 100 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 99997 | 99998 | Male | 66.0 | 0 | 0 | 0.207983 | 5.7 | 155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 99998 | 99999 | Female | 24.0 | 0 | 0 | 0.296569 | 4.0 | 100 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 99999 | 100000 | Female | 57.0 | 0 | 0 | 0.144958 | 6.6 | 90 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

[100000 rows x 15 columns]

**Figure 5 MapReduce-based Pre-processed results**

As shown in the above figure, Batch Normalized Covariate Transpose Propagated-based Pre-processing applies Delta One-hot Encoding and Batch Normalized Covariate Transpose-based Normalization via Map Reduce mechanism. By applying above two functions separately, mapper task performs process parallel. Next, using intermediate key-value pairs generated for both Smoking_History feature and BMI feature are shuffled and pre-processed results are sent to reducer tasks. When applying this pre-processed sample results for diagnosing stage in turn minimizes the overall training process. Following which with the pre-processed sample results

relevant features is selected using Non-linear Auto Correlated Encoding Feature Selection algorithm. By applying this algorithm first, highly linear correlated feature were selected and then features though found to be relevant however non-linear are selected as shown in figure 6.

**Non-linear Auto Correlated Encoding Feature Selected Results**

```
Correlation Coefficient Values:
diabetes                1.000000
blood_glucose_level     0.419558
HbA1c_level             0.400660
age                     0.258008
hypertension            0.197823
heart_disease           0.171727
bmi                     0.162222
smoking_history         0.152345
gender                  0.048210

 correlation coefficient of random features = ['blood_glucose_level', 'HbA1c_level', 'age', 'hypertension', 'heart_disease', 'smoking_history']

 After Non-linear autoencoder encoding = ['blood_glucose_level', 'HbA1c_level', 'age', 'hypertension', 'heart_disease', 'bmi', 'smoking_history']

 Final Selected Features = ['blood_glucose_level', 'HbA1c_level', 'age', 'hypertension', 'heart_disease', 'bmi', 'smoking_history']
```

✅ Final Selected Features:

| | ID | blood_glucose_level | HbA1c_level | age | hypertension | heart_disease | bmi | smoking_history_never |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 140 | 6.6 | 80.0 | 0 | 1 | 25.19 | 1 |
| 1 | 2 | 80 | 6.6 | 54.0 | 0 | 0 | 27.32 | 0 |
| 2 | 3 | 158 | 5.7 | 28.0 | 0 | 0 | 27.32 | 1 |
| 3 | 4 | 155 | 5.0 | 36.0 | 0 | 0 | 23.45 | 0 |
| 4 | 5 | 155 | 4.8 | 76.0 | 1 | 1 | 20.14 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 99995 | 99996 | 90 | 6.2 | 80.0 | 0 | 0 | 27.32 | 0 |
| 99996 | 99997 | 100 | 6.5 | 2.0 | 0 | 0 | 17.37 | 0 |
| 99997 | 99998 | 155 | 5.7 | 66.0 | 0 | 0 | 27.83 | 0 |
| 99998 | 99999 | 100 | 4.0 | 24.0 | 0 | 0 | 35.42 | 1 |
| 99999 | 100000 | 90 | 6.6 | 57.0 | 0 | 0 | 22.43 | 0 |

| smoking_history_No Info | smoking_history_current | smoking_history_former | smoking_history_ever | smoking_history_not current |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | ... | ... | ... | ... |
| ... | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |

[100000 rows x 13 columns]

**Figure 6 Feature selected results**

As shown in the above generated results by employing both linear and non-linear features for selection, misclassification rate is said to be improved with high rate of precision. Finally with the above relevant feature selected results, the diabetes disease detection is performed by applying Normalized Exponential Classifier as shown in figure 7.

**Normalized Exponential Classified Results for diabetic disease prediction**

```
Feature columns: ['ID', 'age', 'HbA1c_level', 'blood_glucose_level', 'diabetes']

Prediction Rule:
if age >= 45 AND HbA1c_level >= 6.5 AND blood_glucose_level >= 126:
    diabetes = 1
else:
    diabetes = 0
```

| | ID | age | HbA1c_level | blood_glucose_level | diabetes | \ |
|---|---|---|---|---|---|---|
| 0 | 1 | 80.0 | 6.6 | 140 | 1 | |
| 1 | 2 | 54.0 | 6.6 | 80 | 1 | |
| 2 | 3 | 28.0 | 5.7 | 158 | 1 | |
| 3 | 4 | 36.0 | 5.0 | 155 | 1 | |
| 4 | 5 | 76.0 | 4.8 | 155 | 1 | |
| ... | ... | ... | ... | ... | ... | ... |
| 99995 | 99996 | 80.0 | 6.2 | 90 | 0 | |
| 99996 | 99997 | 2.0 | 6.5 | 100 | 1 | |
| 99997 | 99998 | 66.0 | 5.7 | 155 | 1 | |
| 99998 | 99999 | 24.0 | 4.0 | 100 | 0 | |
| 99999 | 100000 | 57.0 | 6.6 | 90 | 1 | |

**Figure 7 Classified results for diabetes disease prediction**

As shown in the above results by employing Normalized Exponential function as classifier generates non-negative outputs therefore ensuring that each classified results represents a valid probability. This in turn aids in improving the overall accuracy of diabetic disease detection.

**4.2 Discussion**

In this study, five experiments were conducted to analyze and validate the Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) method along with detailed comparison between two existing methods, Serial Cascaded Convolutional Ensemble Net work (SCCEN) [1] and Deep Belief Network with Generative Adversarial Networks (DBN-GAN) [2] for diabetes prediction Also a detailed quantitative analysis is made for five different performance

metrics, training time, misclassification rate, precision, recall and accuracy.

#### 4.2.1 Performance analysis of training time

In this section the training time involved in diabetic prediction process is analyzed and validated. A considerable amount of time is said to be consumed during diabetic prediction and this is referred to as the training time. The training time is measured as given below.

$$TT = \sum_{i=1}^{m} \quad S_i * Time\ (Res_i) \tag{13}$$

From the above equation (13) training time '$TT$' is measured based on the samples involved in the simulation process '$S_i$' and the time consumed in generating the results '$Time\ (Res_i)$' as either diabetic or non-diabetic. It is measured in terms of seconds (sec). Table presents the performance metrics of training time for diabetes prediction.

**Table Training time using NACE-NEC, SCCEN [1] and DBN-GAN [2]**

| Samples | Training time (sec) | | |
|---|---|---|---|
| | NACE-NEC | SCCEN [1] | DBN-GAN [2] |
| 8000 | 280 | 344 | 384 |
| 16000 | 315 | 385 | 405 |
| 24000 | 335 | 390 | 425 |
| 32000 | 385 | 430 | 455 |
| 40000 | 425 | 460 | 475 |
| 48000 | 455 | 500 | 505 |
| 56000 | 485 | 540 | 535 |
| 64000 | 515 | 570 | 585 |
| 72000 | 535 | 590 | 625 |
| 80000 | 555 | 615 | 645 |

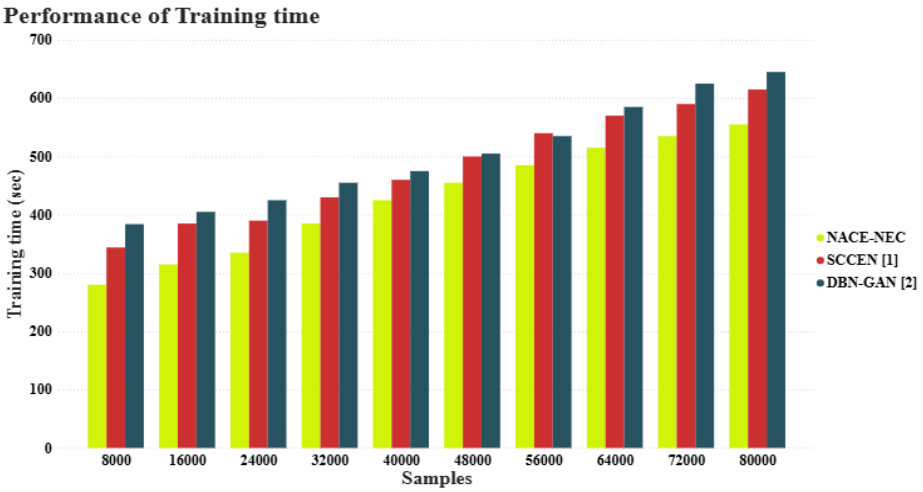


**Figure 8 Diagrammatic representation of training time**

Figure 8 shows the graphical representation of training time using the proposed NACE-NEC and two existing methods, SCCEN [1], DBN-GAN [2] for diabetic disease prediction. While horizontal axis marks the sample instances involved in simulation for the three methods, its training time results are obtained by substituting the values in equation 13. An upsurge in training time is obtained using all the three methods, however, minimal using NACE-NEC method comparing to [1] and [2], therefore improving the efficiency of the proposed method. The reason was that by applying Batch Normalized Covariate Transpose Propagated MapReduce as pre-processing first performed Delta One-hot Encoding for Smoking_History feature and then Batch Normalized Covariate Transpose for BMI feature. These results were then applied via Map Reduce mechanism. Following which using intermediate key-value pairs generated for both Smoking_History feature and BMI feature shuffling process was performed and grouped via ID. This in turn minimized the overall training process involved in diabetic prediction using proposed NACE-NEC method by 13% compared to [1] and 19% compared to [2].

#### 4.2.2 Performance analysis of misclassification rate

Misclassification rate is a performance metric that quantifies the ratio of incorrect predictions made by a classification model. Misclassification rate evaluates the frequency at which a method allocates wrong label to sample instances. A lower misclassification rate denotes better method performance, representing that the method is more accurate in its diabetic predictions. The misclassification rate is mathematically stated as given below.

$$MR = \sum_{i=1}^{m} \quad \frac{S_{IP}}{S_i} * 100 \tag{14}$$

From the above equation (14) misclassification rate '$MR$' is measured based on the number of incorrect predictions made '$S_{IP}$' and the total number of predictions '$S_i$'. It is measured in terms of percentage (%).Table presents the performance metrics of misclassification rate for diabetes prediction.

**Table Misclassification rate using NACE-NEC, SCCEN [1] and DBN-GAN [2]**

| Samples | Misclassification rate (%) | | |
|---|---|---|---|
| | NACE-NEC | SCCEN [1] | DBN-GAN [2] |
| 8000 | 0.93 | 1.12 | 1.43 |
| 16000 | 1.08 | 1.28 | 1.43 |
| 24000 | 1.25 | 1.45 | 1.6 |
| 32000 | 1.55 | 1.75 | 1.9 |
| 40000 | 1.95 | 2.15 | 2.3 |
| 48000 | 2.15 | 2.35 | 2.5 |
| 56000 | 1.85 | 2.05 | 2.2 |
| 64000 | 1.65 | 1.85 | 2 |
| 72000 | 1.35 | 1.55 | 1.7 |
| 80000 | 1.55 | 1.75 | 1.9 |

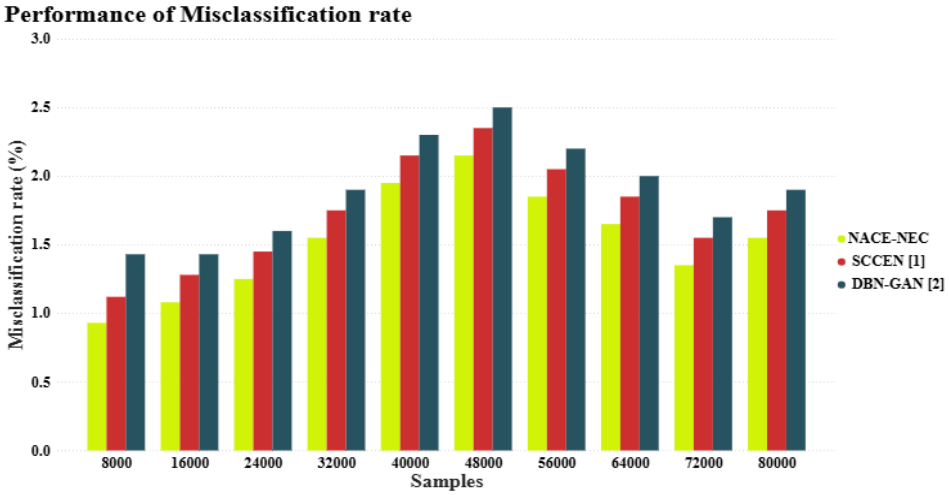


**Figure 9 Diagrammatic representation of misclassification rate**

Figure 9 given above shows the graphical representation of misclassification rate using the three methods, NACE-NEC, SCCEN [1] and DBN-GAN [2]. While performing the classification process by differentiating between samples as diabetic and non-diabetic certain amount of misclassification occurs. From the above graphical representation the misclassification rate was observed to be higher when samples were increased for the first six iterations however for the next three iterations the misclassification rate was reduced. This was evident for all the three methods, however comparative analysis showed minimum amount of misclassification rate using NACE-NEC method upon comparison to [1] and [2]. The reason was that by applying the Non-linear Auto Correlated Encoding for selecting relevant features for diagnosis two different processes were performed, correlation analysis and feature importance ranking. By performing these two functions separately the most significant feature were analyzed via correlation coefficient function. Following which single set of significant features were arrived at based on the feature importance ranking employing non-linear auto encoding. This in turn aided in minimizing the misclassification rate of NACE-NEC method by 14% compared to [1] and 26% compared to [2].

#### 4.2.3 Performance analysis of precision, recall and accuracy

In the classification process, the prediction can be one of four special cases. If the actual value of the target in the diabetes prediction dataset is true and the method employed predicts it as such, then the prediction is a True Positive (TP). On the other hand, if the classifier predicts it as False, then the prediction is said to be False Negative (FN). In a similar manner, if the actual value of the target in the dataset is False and the proposed method predicts it as such, then the prediction is True Negative (TN). On the contrary, if the classifier predicts it as true then the prediction is said to be false positive. Precision is the percentage of samples that a classifier has labeled as positive with respect to the total predictive positives. This is mathematically formulated as given below.

$$Pre = \frac{TP}{TP+FP} \quad (15)$$

From the above equation () precision '$Pre$', is measured by utilizing the true positive '$TP$' (i.e. diabetic patient identified as diabetic) and false positive '$FP$' (i.e. non-diabetic patient identified as non-diabetic). Recall gives information about the percentage of True Positives that is correctly classified during the test. This is mathematically represented as given below.

$$Rec = \frac{TP}{TP+FN} \quad (16)$$

From the above equation (16) recall '$Rec$' is measured using the true positive '$TP$' and false negative '$FN$'. Accuracy on the other hand is the percentage of the correct predictions that that a classifier (i.e. method in comparison) has made compared with actual values of the

target in testing. This is mathematically stated as given below.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

From the above equation () accuracy '$Acc$' is evaluated based on the true positive '$TP$', true negative '$TN$', false positive '$FP$' (i.e. diabetic patient identified as non-diabetic) and false negative '$FN$' respectively (i.e. non-diabetic patient identified as diabetic). Table presents the performance metrics of precision, recall and accuracy for diabetes prediction.

**Table Precision, Recall and Accuracy using NACE-NEC, SCCEN [1], DBN-GAN [2]**

| Samples | Precision | | | Recall | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|
| | NACE-NEC | SCCEN [1] | DBN-GAN [2] | NACE-NEC | SCCEN [1] | DBN-GAN [2] | NACE-NEC | SCCEN [1] | DBN-GAN [2] |
| 8000 | 0.98 | 0.98 | 0.97 | 0.99 | 0.98 | 0.98 | 0.98 | 0.97 | 0.96 |
| 16000 | 0.95 | 0.88 | 0.84 | 0.97 | 0.87 | 0.82 | 0.96 | 0.88 | 0.83 |
| 24000 | 0.92 | 0.85 | 0.81 | 0.93 | 0.83 | 0.78 | 0.93 | 0.85 | 0.8 |
| 32000 | 0.89 | 0.82 | 0.78 | 0.9 | 0.8 | 0.75 | 0.91 | 0.83 | 0.78 |
| 40000 | 0.85 | 0.78 | 0.74 | 0.87 | 0.77 | 0.72 | 0.87 | 0.79 | 0.74 |
| 48000 | 0.81 | 0.74 | 0.7 | 0.89 | 0.79 | 0.74 | 0.83 | 0.75 | 0.7 |
| 56000 | 0.83 | 0.76 | 0.72 | 0.92 | 0.82 | 0.77 | 0.85 | 0.77 | 0.72 |
| 64000 | 0.86 | 0.79 | 0.75 | 0.94 | 0.84 | 0.79 | 0.88 | 0.8 | 0.75 |
| 72000 | 0.89 | 0.82 | 0.78 | 0.97 | 0.87 | 0.82 | 0.9 | 0.82 | 0.77 |
| 80000 | 0.92 | 0.85 | 0.81 | 0.96 | 0.86 | 0.81 | 0.92 | 0.83 | 0.78 |

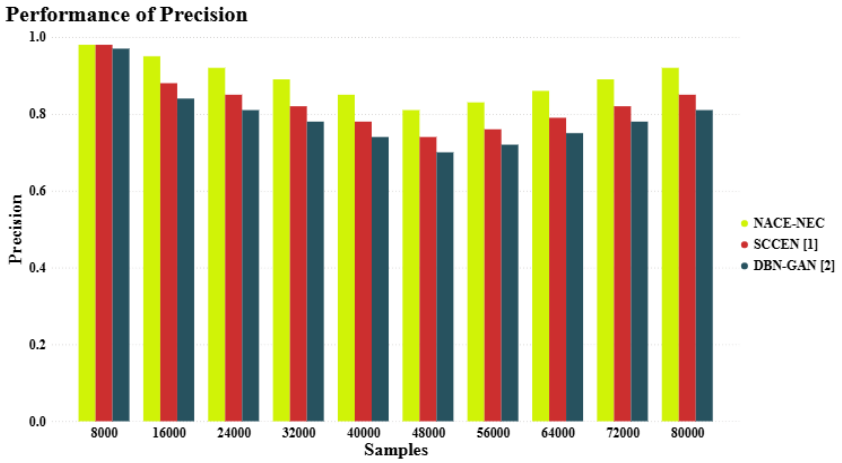


**Figure 10 Diagrammatic representation of precision**

Figure 10 given above illustrate the precision results using three methods, NACE-NEC, SCCEN [1] and DBN-GAN [2]. In order to ensure fair comparisons, similar sample instance from same dataset of each ID is used and results are analyzed. The results from above graph show better precision results using proposed NACE-NEC method compared to [1] and [2].

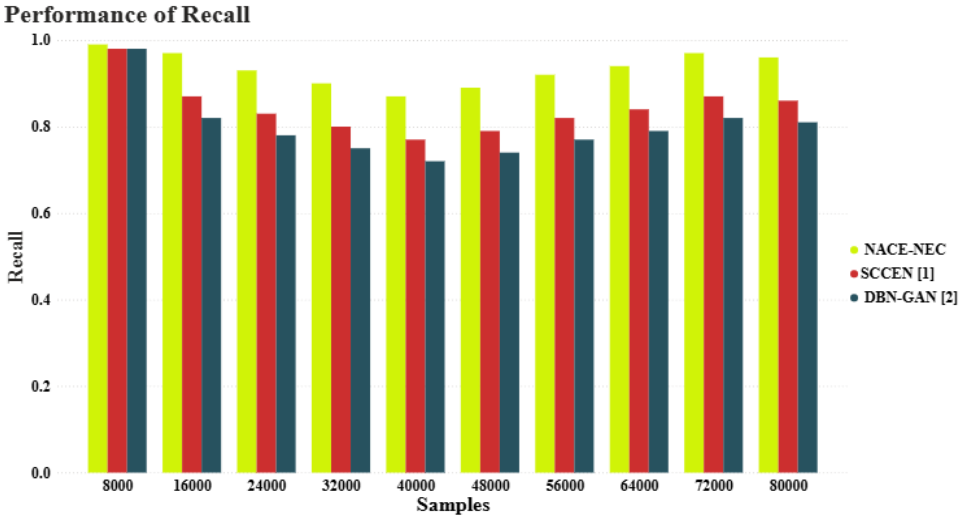


**Figure 11 Diagrammatic representation of recall**

Figure 11 given above presents the results of recall rate involved in diabetic detection process using the proposed NACE-NEC and two existing methods, SCCEN [1] and DBN-GAN. To obtain the recall rate results, the values of true positive and false negative using the three methods were validated and analyzed. With the validated value, recall rate using the three methods was determined by substituting it in the equation (16). The graphical representation show better recall results using proposed NACE-NEC method when compared to [1] and [2].

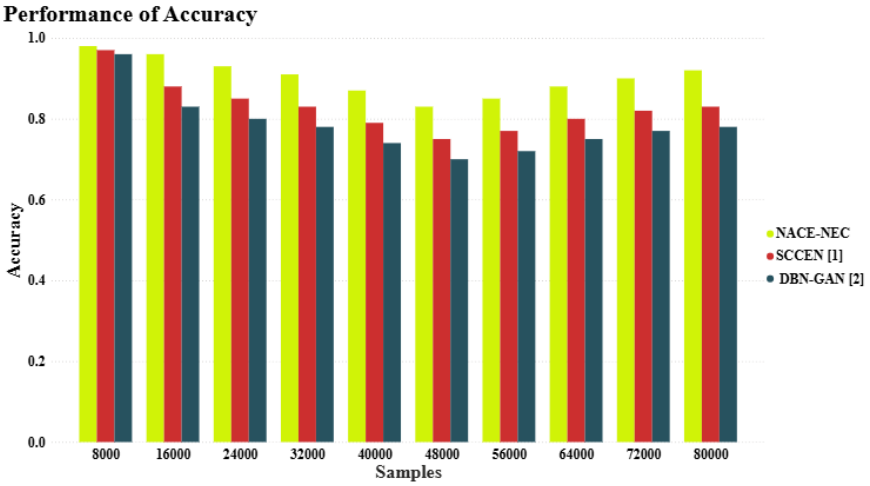


**Figure 12 Diagrammatic representation of accuracy**

Finally figure 12 given above plot accuracy performance metric using three different methods namely NACE-NEC, SCCEN [1] and DBN-GAN for diabetes disease detection Here, in addition to the values of true positive and false negative, true negative and false positive results are measured and the values were substituting in equation (17) to generate accuracy results. From the above graphical representation, accuracy results of NACE-NEC method were found to be better than [1] and [2].

The precision improvement using proposed NACE-NEC method was owing to the application of Non-linear Auto Correlated Encoding Feature Selection algorithm via the first hidden layer. Here, by applying this algorithm, highly correlated features using correlation coefficient function was applied to obtain the most significant feature. Following which along with the most significant feature obtained via correlation coefficient, non-linear auto encoding function was applied to ascertain the non-linear nature of feature. Accordingly based on the feature importance significant features were selected. This in turn selected the features that were strongly associated to the target variable, even if the association was not found to be linear, therefore improving precision using proposed NACE-NEC method by 7% compared to [1] and 11% compared to [2]. Second the recall improvement using proposed NACE-NEC method would be contributed to the application of back propagation function based feature selection algorithm via the second hidden layer. By applying this algorithm the evaluated results were compared using a loss function to measure the error. This in turn propagated backward to fine-tune the weights iteratively until error rate reaches an acceptable level. This in turn minimized the false negative rate and improving overall recall using proposed NACE-NEC method by 10% compared to [1] and 15% compared to [2]. Third the accuracy improvement using proposed NACE-NEC method was due to the application of Normalized Exponential Classification for diabetic disease prediction via the third hidden layer. By applying this

classifier convert vector of raw scores into a probability distribution, therefore producing non-negative outputs that sum to one, ensuring that each classified results represents a valid probability. Owing to this precise and accurate differentiation between classes (i.e. diabetic and non-diabetic) is made in a significant manner. This in turn improves overall accuracy of proposed NACE-NEC method by 8% compared to [1] and 13% compared to [2].

## 5. CONCLUSION

This paper presented a novel deep learning method using Non-linear Auto Correlated Encoding and Normalized Exponential Classification (NACE-NEC) for detecting diabetes disease. The proposed NACE-NEC method performed via deep learning performs the tasks of pre-processing, feature extraction and classification for diabetes disease prediction via one input layer, three hidden layers and one output layer. With the collected data, Batch Normalized Covariate Transpose Propagated MapReduce-based Pre-processing (first hidden layer) was applied to generate normalized results for further processing. Then, the most relevant features for differentiating between diabetic and non-diabetic were selected (second hidden layer) using non-linear auto encoding-based feature importance ranking. Next, the actual classification process using Normalized Exponential Classification for diabetes detection was performed. The algorithm is outlined to optimize five performance metrics, reducing training time, misclassification rate, with improved precision, recall and accuracy. The proposed NACE-NEC method is assessed on various numbers of samples. The performance of the proposed NACE-NEC method is compared to the popular diabetic disease detection methods such as SCCEN and DBN-GAN [2] using different performance metrics. Experimental results prove that the NACE-NEC method efficiently identifies and differentiates between diabetic and non-diabetic significantly.

## REFERENCES


[1] Santosh Kumar Bejugam, Jyothi Vankara, “An efficient model for diabetic detection using heuristic approach based serial cascaded convolutional ensemble network”, Artificial Intelligence Review, Aug 2025 [Serial Cascaded Convolutional Ensemble Network (SCCEN)]

[2] Olusola Olabanjo, Ashiribo Wusu, Olufemi Olabanjo, Mauton Asokere, Oseni Afisi, Boluwaji Akinnuwesi, “A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly imbalanced data”, International Journal of Information Technology, Springer, Vol. 17, Mar 2025 [Deep Belief Network with Generative Adversarial Networks (DBN-GAN)]

[3] Mohd Usama, Belal Ahmad, Jiafu Wan, M. Shamim Hossain, Mohammed F. Alhamid and M. Anwar Hossain, “Deep Feature Learning for Disease Risk Assessment based on Convolutional Neural Network with Intra-layer Recurrent Connection by using Hospital Big Data”, IEEE Access, Apr 2018

[4] John Martinsson, Alexander Schliep, Bjorn Eliasson, Olof Mogren, “Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks”, Journal of Healthcare Informatics Research, Springer, Dec 2019

[5] Mehmet Akif Bulbu, “A novel hybrid deep learning model for early stage diabetes risk prediction”, The Journal of Supercomputing, Springer, May 2024

[6] Melih Agraz1 Ero, Egrioglu, Eren Baş, Mu-Yen Chen, Dinçer Goksuluk, Mehmet Furkan Burak, “Diabetes Development Prediction Using a Hybrid Model Combining Dendritic Artificial Neuron Model and Logistic Regression”, Endocrinology, Research and Practice, Mar 2025

[7] Hung Viet Nguyen, Younsung Choi, Haewon Byeon, “Anexplainable hybrid deep learning model for prediabetes prediction in men aged 30 and above”, Journal of Men’s Health, Vol. 20, Oct 2024

[8] Hayato Tanabe, Masahiro Sato, Akimitsu Miyake, Yoshinori Shimajiri, Takafumi Ojima, Akira Narita, Haruka Saito, Kenichi Tanaka, Hiroaki Masuzaki, Hideki Katagiri, Gen Tamiya, Eiryo Kawakami, Michio Shimabukuro, “Machine learning-based reproducible prediction of type 2 diabetes subtypes”, Diabetologia, Springer, Aug 2024

[9] Boon Feng Wee, Saaveethya Sivakumar, King Hann Lim, W. K. Wong, Filbert H. Juwono, “Diabetes detection based on machinelearning anddeep learning approaches”, Multimedia Tools and Applications, Springer, Aug 2023

[10] Kiran Kumar Patro, Jaya Prakash Allam, Umamaheswararao Sanapala, Chaitanya Kumar Marpu, Nagwan Abdel Samee, Maali Alabdulhafith, Pawel Plawiak, “An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques”, BMC Bioinformatics, Oct 2023

[11] Huaping Zhou, Raushan Myrzashova, Rui Zheng, “Diabetes prediction model based on an enhanced deep neural network”, EURASIP Journal on Wireless Communications and Networking, Nov 2020

[12] Abeer El-Sayyid El-Bashbishy, Hazem M. El-Bakry, “Pediatric diabetes prediction using deep learning”, Scientific Reports, Mar 2024

[13] Rasool Esmaeilyfard, Mohsen Bayati, “Enhancing AI-driven forecasting of diabetes burden: a comparative analysis of deep learning and statistical models”, Scientific Reports, Mar 2025

[14] Neha Katiyar, Hardeo Kumar Thakur, Anindya Ghatak, “Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review”, e-Prime - Advances in Electrical Engineering, Electronics and Energy, Elsevier, Vol. 9, Sep 2024

[15] Jobeda Jamal Khanam, Simon Y. Foo, “A comparison of machine learning algorithms for diabetes prediction”, The Korean Institute of Communications and Information Sciences, Feb 2021

[16] Nor Nisha Nadhira Nazirun Ali Selamat, Asnida Abdul Wahab, Hamido Fujita, Ondrej Krehcar, Kamil Kuca, Ganhongseng, “Prediction Models for Type 2 Diabetes Progression: A Systematic Review”, IEEE Access, Vol. 12, Nov 2024

[17] Toshita Sharma, Manan Shah, “A comprehensive review of machine learning techniques on diabetes detection”, Visual Computing for Industry, Biomedicine, and Art, Springer, Nov 2021

[18] R. Ramani, S. Edwin Raja, D. Dhinakaran, S. Jagan, G. Prabaharan, “MapReduce based big data framework using associative Kruskal poly Kernel classifier for diabetic disease prediction”, MethodsX, Elsevier, Vol. 14, Jun 2025

[19] Kannadasan K, Damodar Reddy Edla, Venkatanareshbabu Kuppili, “Type 2 diabetes data classification using stacked autoencoders in deep neural networks”, Clinical Epidemiology and Global Health, Elsevier, Vol. 7, Dec 2019

[20] Huma Naz, Sachin Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset”, Journal of Diabetes & Metabolic Disorders, Springer, Apr 2020

[21] Jian Zhao, Hanlin Gao, Lei Sun, Lijuan Shi, Zhejun Kuang, Haiyan Wang, “Type 2 diabetes prediction method based on dual-teacher knowledge distillation and feature enhancement”, Scientific Reports, May 2025

[22] Dr. Sheetalrani R Kawale, Pooja Kallappagol, Dr. Sharankumar Ho, "Deep Neural Network Approach for Early-Stage Diabetes Risk Prediction using Hybrid SMOTE-ENN and GAN with SHAP-Based Feature Explanations", Journal of Neonatal Surgery, Vol. 14, Jan 2025

[23] Tawfik Beghriche, Mohamed Djerioui, Youcef Brik, Bilal Attallah, Samir Brahim Belhaouari, "An Efficient Prediction System for Diabetes Disease Based onDeep Neural Network", Complexity, Wiley, Dec 2021

[24] Jayakumar Kaliappan, J Saravana Kumar, S Sundaravelan, T Anesh, R Rithik Yashbir Singh, Diana V Vergarcia, Yassine Himeur, Wathiq Mansoor, Shadi Atalla, Kathiravan Srinivasan, " Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets", Frontiers in Artificial Intelligence, Vol. 7, Aug 2024

[25] Wenguang Li, Yan Peng, Ke Peng, "Diabetes prediction model based on GA XGBoost andstacking ensemble algorithm", PLOSONE|, Sep 2024

[26] Tuan Minh Le, Thanminhvo, Tan Nhat Pham, Sonvutruong Dao, "A Novel Wrapper Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic", IEEE Access, Vol. 9, Jan 2021

[27] Muhammad Mazhar Bukhari, Bader Fahad Alkhamees Abdu Gumaei, Adel Assiri, Syed Sajid Ullah, Saddam Hussain, "An Improved Artificial Neural Network Model for Effective Diabetes Prediction", Complexity, Wiley, Apr 2021

[28] Xiaobo Qi, Yachen Lu, Ying Shi, Hui Qi, Lifang Ren, "A deep neural network prediction method for diabetes based on Kendall's correlation coefficient and attention mechanism", PLOSONE, Jul 2024

[29] Iqra Naveed, Muhammad Farhat Kaleem, Karim Keshavjee, Aziz Guergachi, "Artificial intelligence with temporal features outperforms machine learning in predicting diabetes", PLOS Digital Health, Oct 2023

[30] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

 