Received: February 28, 2024. Revised: March 23, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Diabetes Prediction and Classification Using Self-adaptive Evolutionary Algorithm with Convolutional Neural Network

Suhas Kamshetty Chinnababu[1]* **Anandababu Jayachandra[1]**

[1]*Department of Information Science and Engineering, Malnad College of Engineering, Hassan and Visvesvaraya Technological University, Belagavi, India*
* Corresponding author's Email: suhas2385@gmail.com

Abstract: Globally, diabetes mellitus is the most dangerous disease and it is important to predict disease at an early stage to treat the disease. The learning-based algorithms play a significant part in supporting decision-making in diabetes prediction as well as diagnosis. Machine Learning (ML) based diabetes classification suffers from poor performance due to constraints such as limited labeled data and the challenge of data imbalance. Therefore, this research proposes the Self-adaptive Evolutionary Algorithm (SAEA) with Convolutional Neural Network (CNN) for the prediction and classification of diabetes. For validating the proposed method's effectiveness, the data is collected from benchmark datasets such as the Pima Indian Diabetes dataset (PIDD), Frankfurt Hospital, Germany, and North California State University (NCSU). Then, the min-max normalization is used for normalizing the data in the pre-processing step. The Chi-square-based feature selection technique is used for selecting the best feature for categorial features and finally, the SAEA with CNN is used for classification and it classifies the disease into diabetes and non-diabetes. The proposed method's effectiveness is estimated by using various matrices and it attains the accuracy of 99.87% and 99.99% by using PIDD and German datasets when compared to existing approaches like Deep Neural network (DNN), CNN, and Deep CNN (DCNN).

Keywords: Chi-square, Convolutional neural network, Diabetes mellitus, Machine learning, Self-adaptive evolutionary algorithm.

1. Introduction

Globally, diabetes (diabetes mellitus) is one of the most significant predominant chronic diseases, and it occurs when the glucose in the blood is too high. Diabetes is fatal or extremely lowers the quality of life of both men and women [1]. Diabetes is a chronic pathology that happens only when the glucose level in the blood is excessive. Glucose is the energy source of the body and insulin is the hormone, secreted through the pancreas, which controls the glucose level in the cells to be utilized for energy [2]. There are various types of diabetes such as type 1, type 2 as well as gestational. In type 1, the pancreas generates insulin-dependent diseases, which occur mostly in young people of age less than 30. Type 2 is reasoned through insulin confrontation, which occurs in middle-aged and elder people and is integrated with hypertension, obesity, atherosclerosis, and so on. The gestational is hyperglycemia which happens during pregnancy [3-5]. Diabetes can cause long-term damage to various organs like the heart, blood vessels, eyes as well and kidneys, however, it can be cured by diagnosis and prediction at an early stage. The diagnosis of diabetes is a significant and challenging task in the medical field. Hence, diabetes diagnosis and prediction are significant provocative subjects for the research. For an early diagnosis of this disease, various parameters like plasma glucose concentration, serum insulin, age, blood pressure, and so on have been collected. The traditional-based diabetes prediction requires a prolonged time to analyze and to the final decision [6, 7].

Many algorithms and techniques have been introduced for an application in extracting the understanding as well as data in the diagnosis the diabetes from benchmark datasets [8]. Hence, the

existing researchers used advanced computer and information technologies for early diagnosis of diabetes mellitus [9]. Machine Learning (ML) algorithms like Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes (NB), and so on are used for diabetes prediction instead of traditional algorithms [10-12]. However, ML approaches have some limitations in accuracy as well as a selection of features. To address these limitations, the Deep Learning (DL) algorithms are mainly utilized for the prediction process in medical applications. Various researchers have proved that the DL approaches provide better accuracy results, reducing the error rates. The DL approaches can effectively manage a large number of datasets and can solve complex problems [13, 14]. However, some DL based diabetic predictions have not correctly classified all the input images, due to the images having huge matrices with large numbers of pixels and are complex [15]. Hence, this research proposes the metaheuristic optimization of Self-Adaptive Evolutionary Algorithm (SAEA) with a Convolutional Neural Network (CNN) for diabetic prediction and classification. The primary contributions of this research are listed as follows:

- The Chi-square feature selection technique is utilized for selecting the best feature for categorial features and to maximize the performance.
- This research utilized the SAEA with CNN algorithm for diabetes prediction and classification on various benchmark datasets to enhance the overall performance.
- The proposed method's effectiveness is calculated by using various assessment metrics like accuracy, precision, recall, and F1-score.

This research paper is provided as follows: Section 2 provides the Literature survey. Section 3 presented the proposed methodology. Section 4 gives the results, discussion, and the conclusion of this paper is given in Section 5.

2. Literature survey

In this section, some closely related works of diabetes prediction are described. Recent literature has generated a vast amount of research to predict and classify diabetic patients based on symptoms using DL approaches.

Tawfik Beghriche [16] presented the Deep Neural network (DNN) for an efficient medical decision system for the prediction of diabetes. The existing algorithms in computer vision, language processing, as well as analysis of an image, were performed for the diagnosis and prediction of diabetes. Furthermore, these purposes could combine with medical understanding to enhance the effectiveness, adaptability, and transparency of the decision-making process. The effectiveness of the DNN was compared with the existing ML approaches. However, the Logistic Regression (LR) had attained poor performance in overall metrices and lack of interpretability of their outcomes.

Mahendra Kumar Gourisaria [17] developed the ML approaches for the diagnosis and prediction of diabetes mellitus. The suggested approach was predominantly considered the two-benchmark dataset. ML approaches like SVM, NB, as well as Random Forest, were executed for the classification of diabetic from non-diabetic patients. Afterward, the effectiveness of the suggested method was estimated by minimizing the dimensionality through Linear Discriminant Analysis as well as principal Component Analysis. However, this method could not be spatially invariant to the input data and a large number of training data was needed to obtain better results.

María Teresa García-Ordás [18] introduced the pipeline-based DL approaches for diabetic prediction. The Variational Autoencoder (VAE) as well as Sparse Autoencoder (SAE) were utilized for data and feature augmentation. The CNN was used for the classification of diabetes. The SAE had been trained with CNN, which permitted to obtaining of feedback from every layer in backpropagation to enhance the generated feature quality based on spatial depiction forced through CNN. The CNN minimized the computational time; however, Multilayer Perceptron (MLP) had modified the feature extraction, but enhancement was worse than in convolutional net.

Suja A. Alex [19] presented the effective prediction approach of Deep 1D-CNN (DCNN) for diabetes mellitus. The suggested approach utilized Turkey's approach for the detection of missing values. The normalization technique was utilized for expressing the explicit temporal data to enhance the efficiency of the model. Afterward, the SMOTE oversampling technique was utilized to solve the class imbalance problem. Eventually, the DCNN classification was used for prediction as well as estimated by a selective set of estimation depictions. However, the suggested approach's performance was degraded when the dataset had more noise.

Muhammet Fatih Aslan [20] developed the popular CNN approach for the determination of the diabetes diagnosis. Three various classification strategies were applied: Initially, the input image was provided to the ResNet18 as well as ResNet50 CNN approaches. Then, deep features of the ResNet approaches were fused as well as classified with the

SVM. Finally, selected fusion features were classified through the SVM. However, the ResNet approach had susceptible for overfitting as well as limited interpretability.

Kiran Kumar Patro [21] presented the data modelling approach as well as integrated by the DCNN approach for the accurate diabetic prediction and classification. The pre-processing was performed to the solve the problem of data inconsistence, outliers as well as missing values. The presented approach had applied the three data modelling approaches like statistical, relative as well as logic for the prediction of diabetes. However, the only features with the high correlation degree were selected in relative modelling.

R. V. Aswiga [22] developed an efficient framework for an intrinsic difficult diabetes dataset classification. The suggested approach attempted to introduce diabetes disease prediction as well as classification by the utilization of CNN rather than the utilization of traditional approaches. the CNN approach had designed the utilization of benefits of selected significant text features, that were limited in pre-processing layers. Hence, the suggested approach acquired better accuracy and minimized the computational time needed through CNN with fine-tuning. However, the suggested approach required the large amount of labeled data for training as well as computationally intensive.

The limitations of the above-discussed existing TDC approaches include the lack of interpretability, spatially invariant to the input data, a large number of training data was needed to obtain better results, MLP had modified the feature extraction, but enhancement was worse than in convolutional net, performance was degraded when the dataset had more noise, computationally intensive, limited interpretability, susceptible for overfitting and only features with the high correlation degree were selected in relative modelling. To address these limitations, this research proposes the ESGA with CNN for the prediction and classification of diabetes patients into diabetes and non-diabetes.

3. Proposed methodology

This research is widely employed to enhance the prediction and classification of outcomes as well as the accuracy of results. In this research, an Evolutionary Strategy-based Genetic Algorithm (ESGA) with CNN is proposed for the prediction and classification of diabetes. The pre-processing is performed by using min-max normalization before providing input to the feature selection and classification. Then, the pre-processed data can be utilized for the feature selection process for the data to be cleaned, and normalized. The feature selection approach is done by using the Chi-square technique. Then, selected features can be trained using hyperparameter tuning of SAEA with CNN for prediction and classification. Finally, diabetes can be classified into binary results such as normal or not. Fig. 1 shows the flow diagram of the introduced method.

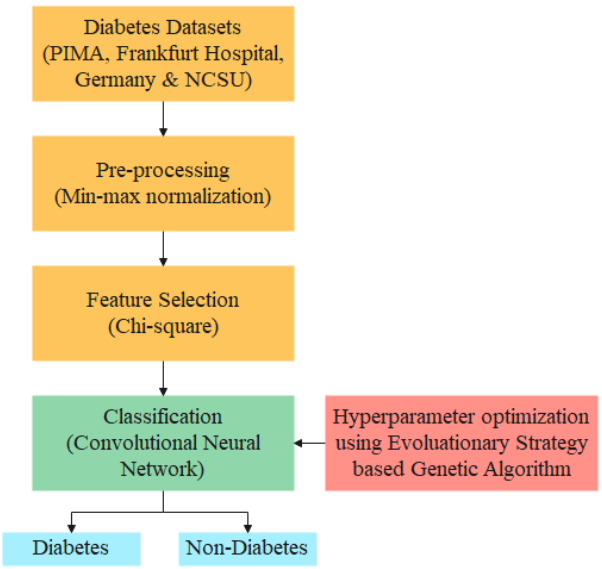


Figure. 1 Flow diagram of the proposed Method

3.1 Dataset collection

In this research, the initial stage of the proposed method is data collection for evaluating the performance of the proposed method. This research utilized the three well-known standard benchmarks such as the Pima Indian diabetes dataset (PIDD) [23], Frankfurt Hospital, Germany [24], and North California State University (NCSU) [25] datasets. The PIDD dataset is the most popular benchmark dataset of diabetes mellitus, which is utilized by various researchers. The PIMA dataset is available in the University of California Irvine (UCI) ML repository. The PIMA dataset contains 9 columns as well as 1 outcome column with the dichotomous value to specify whether the patient has diabetes or not. This dataset contains 768 rows, in which 268 patients have diabetes and 500 are non-diabetes. Also, this dataset contains 9 feature columns such as plasma, glucose, Blood Pressure (BP), pregnancy month, triceps skin fold thickness, patient's age, the function of pedigree, and insulin quality as well as 1 target column (0 or 1) respectively. The value "0" represents that the person is non-diabetic and the value "1" represents that person is diabetic. Frankfurt Hospital, Germany is one of the diabetes benchmark

datasets which involves 9 attributes and 2000 records. In this dataset, 32.4% of the records are diabetes, and the remaining 67.6% are non-diabetes, taking into consideration the fact that all the patients are females aged from 21 to 81. This dataset involves various attributes such as pregnancy frequency, Glucose tolerance, blood pressure, insulin, skin thickness, and diabetes pedigree function, these are described from the range between 0 and 17, 0 to 199, 0 and 122, 0 and 864, 0 and 99 as well as 0.078 to 2.42 respectively. The benchmark NCSU dataset is acquired from NC State University, which involves 442 cases with around 10 attributes. The NCSU feature set is encompassed by age, sex, BMI, and blood glucose level.

These collected datasets are then forwarded to the pre-processing technique for transforming the data into an effective format without any complexities.

3.2 Pre-processing

In this step, the collected diabetes image dataset is utilized as input for pre-processing the data. The pre-processing is significant for transforming the raw data into a useful as well as effective format. The objective of the pre-processing of the image is to accurately extract the features and also to eliminate the noisy data from the fundus image. The pre-processing step supports to minimize the computational complexity of the approach. Generally, the collected input diabetes dataset contains missing values for different attributes, which causes the performance of the proposed methodology. Hence, the min-max normalization is utilized for employing the linear data transformation. A detailed description of this technique is provided as follows:

3.2.1. Min-max normalization

Min-max normalization [26] is a significant step in data pre-processing. This technique normalizes every function by removing its mean and scaling its variance to one, as an identification of a value depends on mean as well as variance. This min-max normalization takes into maximum and minimum values to set the data in the range between 0 and 1 respectively. For each feature, the minimum and maximum value of that feature gets transformed into 0 and 1, which is formulated in Eq. (1) as follows:

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where, $\hat{x}$ – normalized data; x – actual data; x_{min} and x_{max} – minimum and maximum of every feature. This technique will enhance the speed as well as minimize the runtime complexity. The advantages of this technique involve linear, reversible as well as scalable. After that, the pre-processed data can be split into two parts such as 80% of the data being used for training and 20% for testing to avoid misclassification outcomes.

By utilizing the Min-max normalization, the data can be efficiently normalized. Then, this pre-processed output is forwarded to the feature selection process.

3.3 Feature selection

The output of pre-processing is provided as an input for the process of feature selection and it is also known as the selection of attributes. The various existing feature selection approaches concentrate on binary classification problems, which is also known as Recursive feature Elimination (RFE) utilize the classifier to eliminate unimportant features from the dataset. The feature selection reduces the training time as well as eliminates the computational complexity. This process can enhance the model accuracy due to the correct subset selection. In the feature selection process, the Chi-square algorithm is performed, and the detailed description of this technique is described below.

3.3.1. Chi-square

The chi-square [27] is the statistical approach and it is widely utilized for determining whether two variables are independent or not. To maximize the model performance, chi-square is utilized for selecting the best feature from categorial features. The chi-square score is estimated for every pair of feature values as well as class labels. The chi-square test is a majorly utilized statistical approach that can identify whether the two variables are independent or not. For two variables X and Y, the total observed and expected count are estimated. The chi-square is achieved with features f and classes c by utilizing the subsequent Eq. (2) as:

$$\chi^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(s_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (2)$$

Where, s_{ij} – i th feature value along with instances. μ_{ij} – Expected count. This score calculates the amount of deviation between s_{ij} and μ_{ij}. Afterward, the μ_{ij} is formulated in Eq. (3) as:

$$\mu_{ij} = \frac{s_{*j} s_{i*}}{s} \quad (3)$$

Where, s_{i*} - ith value of particular feature; s_{*j} – number of instances in class j ; s – number of instances. The best features can be efficiently selected by using a chi-square and the selected features are then provided to the classification process.

3.4 Classification

The DL algorithms for the detection of diabetics are trained fully and provide a better performance. The deep learning-based classification is widely utilized for the prediction and progression of disease, analyzing as well as categorizing the intrinsic datasets.

3.4.1. Self adaptive evolutionary algorithm

Self-adaptive evolutional Algorithm is the popular stochastic optimization algorithm. It works on the population of solutions as well and every solution is encoded in chromosomes. These solutions are iteratively enhanced by the application of operators consisting of crossover and mutation.

Initialization: An initial population is needed to begin an evolutionary algorithm. Every $X_{i,0} = \{x_{i,0}^1, x_{i,0}^2, \dots x_{i,0}^D\}$ is developed through consistently randomizing the individuals within the search space. Furthermore, initial values of mutation F, as well as crossover Rate CR, are arbitrarily developed with the provided range for every primary discrete actual-value string.

Mutation: Before the mutation is performed, every vector $X_{i,G}$ in the present population is preserved as a target vector. Consistently, the mutation vector $V_{i,G} = \{v_{i,G}^1, v_{i,G}^2, \dots v_{i,G}^D\}$ is developed by enhancing the weight variation among two random vectors to another vector from the present population. The $F_{i,G}$ value of every target vector $X_{i,G}$ is utilized to develop the mutation vector which is expressed in Eq. (4) as:

$$V_{i,G} = X_{a,G} + F_{i,G}(X_{b,G} - X_{c,G}) \tag{4}$$

Where, $X_{a,G}$, $X_{b,G}$ and $X_{c,G}$ – the vectors are arbitrarily chosen from the present population. The directories are arbitrarily developed from every mutation vector.

Crossover: The trial vector $U_{i,G} = \{u_{i,G}^1, u_{i,G}^2, \dots u_{i,G}^D\}$ is generated by choosing the solution component values from $V_{i,G}$ or $X_{i,G}$ utilizing the procedure of crossover, which is equivalent to uniform crossover. Hence, every component within $U_{i,G}$ is expressed in Eq. (5) as:

$$u_{i,G}^j = \begin{cases} v_{i,G}^j, if\ Rand_2 \le CR_{i,G} \\ x_{i,G}^j, \qquad otherwise \end{cases} \tag{5}$$

Where, $u_{i,G}^j$, $v_{i,G}^j$ and $x_{i,G}^j$ - jth parameter in ith trial, mutant as well as target vector. If $Rand_2$ is smaller than $CR_{i,G}$, the value $v_{i,G}^j$ in the mutant is derivative to trial vector, simultaneously, $x_{i,G}^j$ in the target, the vector is copied to the trial vector.

Selection: After crossover, the objective function $f(U_{i,G})$ for every trial vector is estimated. After it is compared with the corresponding $X_{i,G}$ concerning objective function. The vector with minimal objective function survives into further generation $(X_{i,G+1})$ which is expressed in Eq. (6) as:

$$X_{i,G+1} = \begin{cases} U_{i,G}, if\ f(U_{i,G}) \le f(X_{i,G}) \\ U_{i,G}, \qquad otherwise \end{cases} \tag{6}$$

The F and CR values are subjected to the chosen operator. It is an integration of $F_{i,G}$ and $CR_{i,G}$ is capable of generating a better solution $U_{i,G}$ compared to $X_{i,G}$, those two values are provided to $X_{i,G+1}$ and survive to the next generation. The selection of F and CR for the next generation is expressed in Eq. (7) as:

$$F_{i,G+1} = \begin{cases} U_{i,G}, if\ f(U_{i,G}) \le f(X_{i,G}) \\ F_l + Rand_3(F_u - F_l), if\ f(U_{i,G}) > f(X_{i,G}) \end{cases}$$
$$CR_{i,G+1} = \begin{cases} CR_{i,G}, if\ f(U_{i,G}) \le f(X_{i,G}) \\ CR_l + Rand_4(CR_u - CR_l), if\ f(U_{i,G}) > f(X_{i,G}) \end{cases} \tag{7}$$

Where, $Rand_3$, $Rand_4$ – independently developed arbitrary numbers in the range between 0 and 1 respectively. the selected data from the benchmark dataset, for the chi-square technique has been utilized for the identification of the fitness of an agent. Then obtained outcome of the hyperparameter tuning is provided to the CNN algorithm.

3.4.2. Convolutional Neural Network

A CNN is one of the DL approaches and is widely utilized for the classification, regression, and recognition process. The CNN utilizes the bi-dimensional input and is capable of extracting the difficult data features and capable to automatically learning the features. Therefore, this develops the deadly process of an automatic extraction and

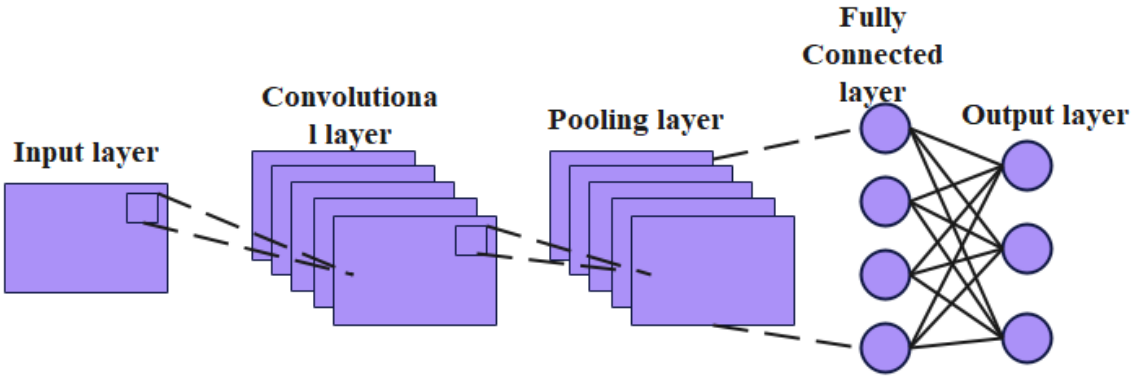


Figure. 2 Basic CNN architecture

description of features in training phases because classification error is reduced. In training, the network alters the filter weights intending to take out an accurate feature map of every class. The distinctive CNN consists of various layers such as input, convolution, max pooling as well as output layers. Fig. 2 shows the basic CNN architecture.

The extracted features from every layer depend on number of kernels and their size. The kernel weights are initialized by utilizing an arbitrary weight, which is trained in the model training course. The convolutional layer is formulated in Eq. (8) as:

$$y_n^k = h\left(b_n^k + \sum w_{x,k}^{n-1} * m_x^n\right) \quad (8)$$

Where, n – a number of layers in the coding network; h - activation function; m_x and y_k – input and output feature map; $w_{n,k}$ – kernel weight; $*$ - utilized for convolutional operation; b_k – bias. Basically, to acquire the mapping feature $r_{i,E}$ from the window of sword embedding vector $e_{i:1+z-1}$ in the input sequence, the convolutional layer applies the matrix-to-matrix performance among window and weight matrix W_f of size to every filter is expressed in Eq. (9) as:

$$r_{i,E(dp)} = \sigma\left(W_f^T \circ e_{i:1+z-1} + b\right), W_f \epsilon R^{dim \times s} \quad (9)$$

Where, $s \in [1.n+m]$; $\sigma(.)$ – nonlinear activation function of convolutional operation; $\circ$ - Hadamard product between two matrices. After the convolutional layer, the outcome is provided to the max-pooling layer. The pooling layers down sampling along the sequential temporal dimension, thereby minimizing the feature map dimensionality, while recollecting the most significant data. The most often type of pooling is Max-pooling, which carries the maximum value in every window. The Fully Connected (FC) Neural Network Layer, in which every neuron is connected to each neuron in the previous layer. The FC layer output through an activation function f is formulated as Eq. (10) as,

$$FC(x) \coloneqq f(Wx = b) \in R^m \quad (10)$$

Every FC layer utilizes the Rectified Linear Unit (ReLU) activation function to develop nonlinearity into model. The ReLU retains the convolutional layer in a specific range. The ReLU is represented as Eq. (11) as:

$$ReLU(x) = \max(0, x)\, x \in R \quad (11)$$

The $ReLU(x) = 1$ for $x > 0$ and $ReLU - (x) = 0$ for $x < 0$. The output of the max-pooling layer into the softmax layer is expressed in Eq. (12) as:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (12)$$

Where z – softmax function input vector contains "n" features of "n" target values; zi - ith item of input vector; e^{z_j} – standard exponential function; $\sum e^{z_j}$ – normalization term to acquire valid probability distribution. Eventually, a softmax layer and the likelihood logarithms are used as input for the examination of diabetes classes. Therefore, diabetes disease is significantly classified into diabetes and non-diabetes by using CNN.

Table 1. Hyperparameter setting of the proposed method

Method	Optimization Algorithm
Population_size	50
Num_generation	10
Max_iteration	100
Learning rate	0.001
Batch size	32

Table 2. Performance analysis of feature selection

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Frankfurt Hospital, Germany	PCC	89.75	89.10	89.50	89.80
	Mutual information	94.90	95.00	94.78	94.85
	Chi-Square	97.89	96.34	98.23	98.89
PIDD	PCC	92.05	92.99	91.86	92.10
	Mutual information	90.56	90.67	90.08	90.45
	Chi-Square	97.67	96.34	97.12	98.35
NCSU	PCC	92.72	92.98	92.53	92.75
	Mutual information	87.69	87.73	87.50	87.64
	Chi-Square	94.23	93.27	93.12	94.12

Table 3 Performance analysis of hyperparameter tuning with classifier

Dataset	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Frankfurt Hospital, Germany	SAEA-DNN	94.90	95.00	94.78	94.85
	SAEA-RNN	83.75	84.10	83.50	83.80
	SAEA-CNN	99.99	99.99	99.99	99.99
PIDD	SAEA-DNN	91.05	91.99	90.85	91.10
	SAEA-RNN	93.83	94.00	93.40	93.70
	SAEA-CNN	99.87	99.82	99.90	99.86
NCSU	SAEA-DNN	85.75	86.00	85.50	85.80
	SAEA-RNN	81.72	81.98	81.53	81.75
	SAEA-CNN	95.00	95.91	95.23	95.10

4. Results and discussion

This section illustrates the result and discussion of the proposed ESGA with CNN for diabetic prediction and classification. The collected three benchmark datasets are used for evaluating the performance of the model. Furthermore, this section represents the experimental setup, evaluation metrics, performance analysis, and comparative analysis. Table 1 shows the hyperparameter settings of the proposed method.

4.1 Experimental results

The proposed EAGA with CNN for diabetic prediction and classification is executed by utilizing the platform of Python 3.9 with Windows 10 OS, and 16GB RAM with intel-i7 processor. The proposed method is analyzed by utilizing various assessment metrics like accuracy, precision, recall, and F1-score. These metrics are formulated in Eqs. (13)-(16) as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

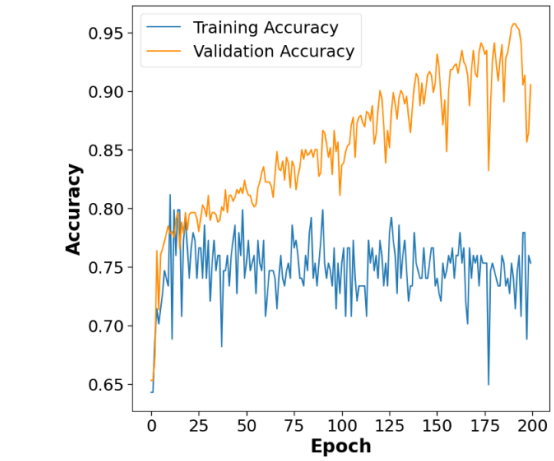
$$F1-score = \frac{2TP}{2TP+FP+FN} \quad (16)$$

Where, TP - True Positive; TN – True Negative; FP – False positive; FN – False Negative.

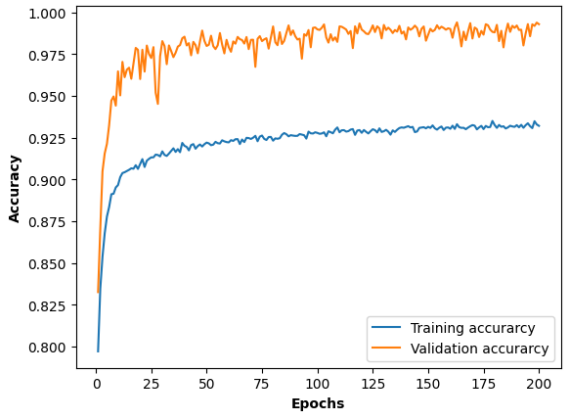
4.2 Performance Analysis

Various existing performance metrics concentrated on the problem of binary classification. To verify the dependability of a proposed SAEA with the CNN model, three different datasets such as PIDD, Germany and NCSU are utilized. The proposed method's effectiveness is determined utilizing accuracy and loss function concerning the number of epochs. Tables 2 and 3 show the performance analysis of feature selection and hyperparameter tuning with the classifier by using various datasets.

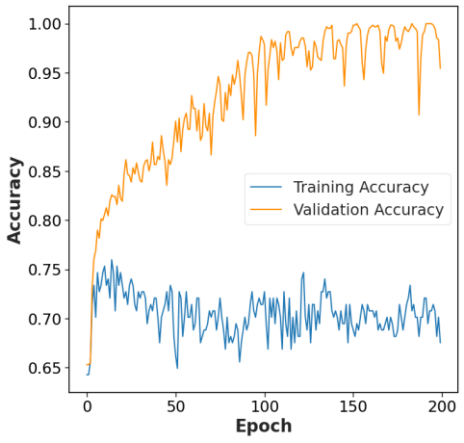
As shown in Table 2, it is obvious that the Chi-square with various feature selection techniques for diabetes prediction by using NCSU, Frankfurt Hospital, Germany and PIDD datasets. The existing feature selection techniques such as Pearson Correlation Coefficient (PCC) and Mutal Information are analyzed. The proposed Chi-square attains the accuracy, precision, recall and F1-score are: 94.23%, 93.27%, 93.12%, and 94.12% in the NCSU dataset, 97.89%. 96.34%, 98.23% and 98.89% in Frankfurt Hospital, Germany dataset as well as 97.67%,



(a)

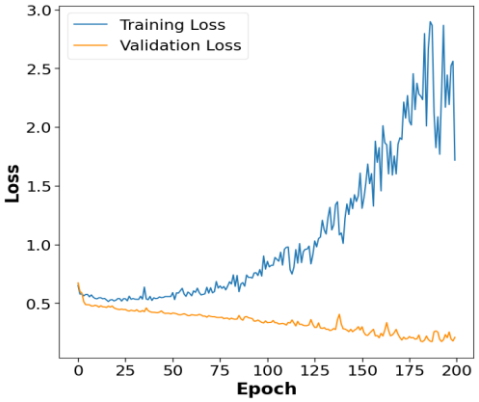


(b)

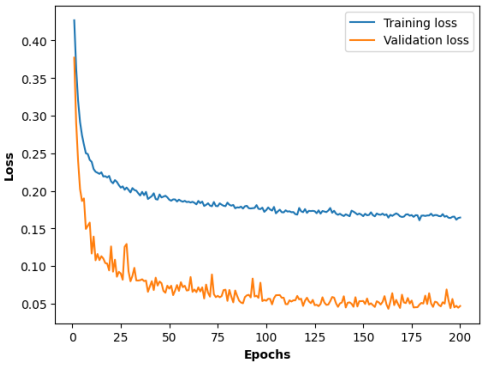


(c)

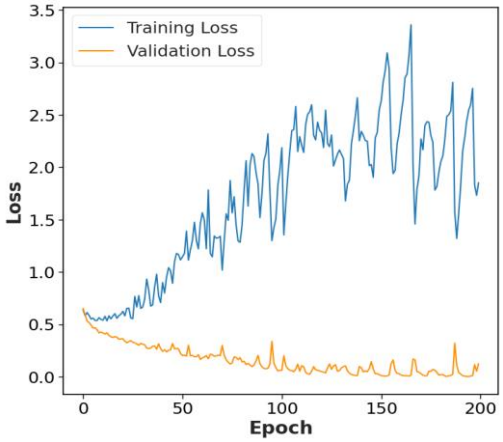
Figure. 3 Graphical representation of accuracy function with various datasets: (a) Frankfurt Hospital, Germany (b) PIDD dataset, and (c) NCSU dataset



(a)



(b)

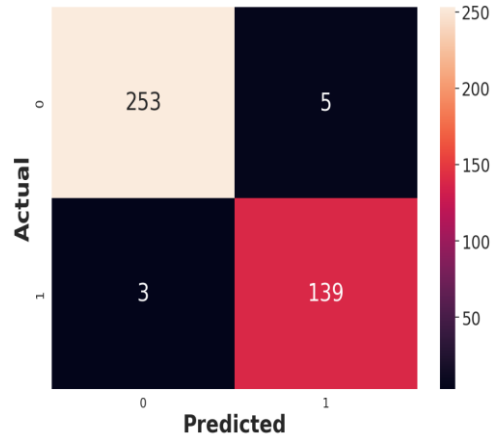


(c)

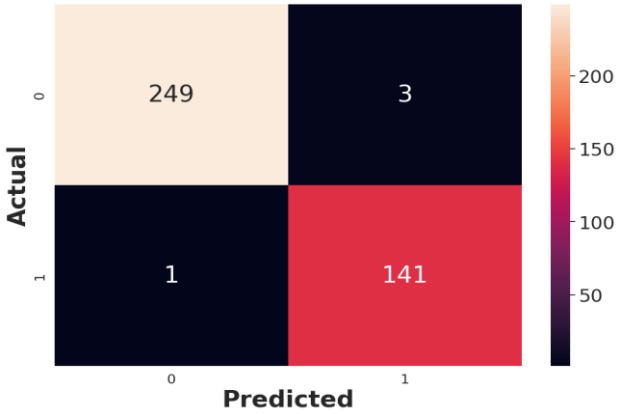
Figure. 4 Graphical representation of loss function with various datasets: (a) Frankfurt Hospital, Germany (b) PIDD dataset, and (c) NCSU dataset

96.34%, 97.12% and 98.35% in the PIDD dataset. The proposed method attains effective outcomes when compared to other competitor approaches with various performance metrics.

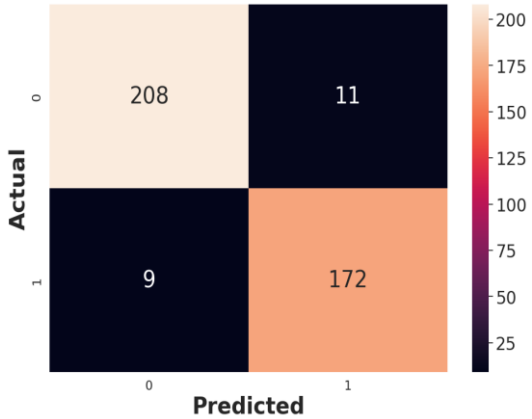
As shown in Table 3, it is obvious that the SAEA with CNN with various classification algorithms for diabetes prediction by using NCSU, Frankfurt Hospital, Germany and PIDD datasets. The existing classifiers such as DNN and Recurrent Neural Networks (RNN) are analyzed. The proposed SAEA with CNN attains the accuracy, precision, recall and F1-score are: 95.00%, 95.91%, 95.23% and 85% in the NCSU dataset, 99.87%, 99.82%, 99.90%, and 99.86% in PIDD dataset. The proposed attains 99.99% in over-performance metrics using Frankfurt Hospital, Germany dataset. The proposed method attains effective outcomes when compared to other



(a)



(b)



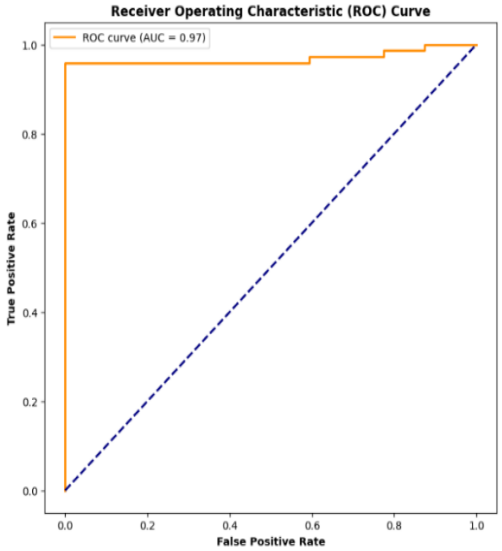
(c)

Figure. 5 Graphical representation of confusion matrix with various datasets: (a) Frankfurt Hospital, Germany (b) PIDD dataset, and (c) NCSU dataset

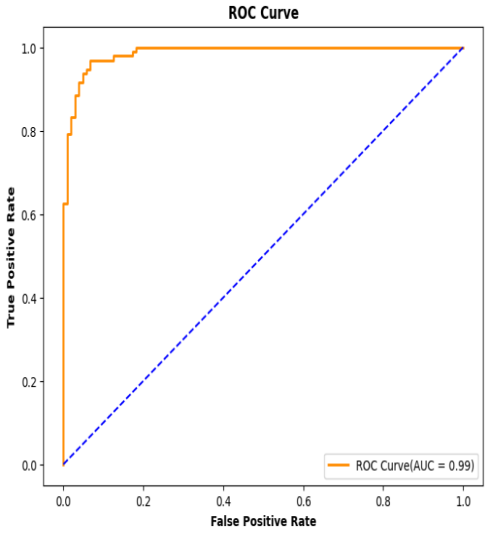
competitor approaches with various performance metrics.

4.2.1. Accuracy function

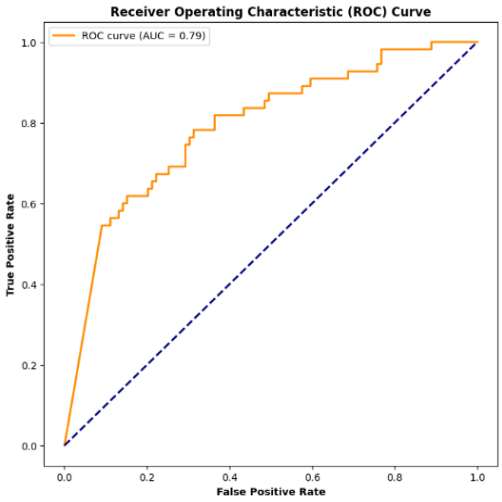
Fig. 3 shows the exponential curve for the accuracy function using various datasets. Fig. 3. (a), (b), and (c) represent the exponential curve for accuracy using Frankfurt Hospital, Germany, PIDD



(a)



(b)



(c)

Figure. 6 Graphical representation of ROC curve with various datasets: (a) Frankfurt Hospital, Germany (b) PIDD dataset, and (c) NCSU dataset

Table 4. Comparative Analysis

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DNN [16]	Frankfurt Hospital, Germany	99.75	N/A	N/A	N/A
ANN [17]	Frankfurt Hospital, Germany	95.80	95.00	95.00	95.00
SAE -CNN [18]	PIDD	92.31	N/A	N/A	N/A
Deep 1D-CNN [19]	PIDD	86.29	N/A	N/A	N/A
ResNet18 and ResNet50-Relief [20]	PIDD	92.19	N/A	N/A	N/A
DCNN with data modelling [21]	PIDD	96.13	94.44	94.42	94.46
Proposed SAEA with CNN	Frankfurt Hospital, Germany	99.99	99.99	99.99	99.99
	PIDD	99.87	99.82	99.90	99.86

and NCSU dataset, which obtains the flat return after an initial rise. The performance of the model is estimated with 200 epochs.

4.2.2. Loss function

Fig. 4 shows the exponential curve for the loss function using various datasets. Fig. 4 (a), (b), and (c) represents the exponentially decreasing curve for loss function with Frankfurt Hospital, Germany, PIDD, and NCSU dataset with respect to number of epochs (200) obtained at the time of validation of a proposed method. The loss function from this dataset obtains minimum loss value with enhancements in epochs.

4.2.3. Confusion matrix

Fig. 5 shows the exponential curve for the confusion matrix using various datasets. Fig. 5 (a), (b), and (c) represents the exponential curve for accuracy using Frankfurt Hospital, Germany, PIDD, and NCSU dataset.

4.2.4. ROC Curve

Fig. 6 shows the ROC curve for various datasets. Fig. 6. (a), (b), and (c) represent the exponential curve for accuracy using Frankfurt Hospital, Germany, PIDD and NCSU dataset, which obtains the flat return after an initial rise. The performance of the proposed method is estimated with True Positive Rate (TPR) with False Positive Rate (FPR).

4.3 Comparative Analysis

Table 4 compares the proposed SAEA with CNN method to a collection of existing algorithms based on various evaluation metrics. The existing methods such as [16-19] are compared by using accuracy, precision, recall, and F1-score respectively.

4.4 Discussion

This research considers the detailed diagnosis and classification of diabetes mellitus by using the various assessment metrics with three benchmark datasets. The proposed SAEA with CNN approach outperforms the baseline as well as competitor approaches are analyzed through the experimentation. The hyperparameter selection values take place a significant role in the effectiveness of DL techniques. Furthermore, the hyperparameter selection of the CNN is dependent on the collected datasets. In comparison analysis, the proposed SAEA with CNN approach is estimated using three benchmark datasets such as PIDD, Frankfurt Hospital, Germany and NCSU datasets achieved better values for the metrics. The proposed method achieved the 99.87%, 99.82%, 99.90%, 99.86% and 99.99%, 99.99%, 99.99%, 99.99% for accuracy, precision, recall, F1-score by using PIDD and Frankfurt Hospital, Germany datasets when compared to the existing methods of ML and DL approaches such as DNN [16], ANN [17], SAE-CNN [18], Deep 1D-CNN [19], ResNet18 and ResNet50-ReliefF [20] and DCNN with data modelling [21] respectively. These results show that the proposed method achieves better results when compared to the existing methods. Eventually, the proposed method shows better performance in overall metrics, furthermore, it successfully carries out the prediction and classification of diabetes.

5. Conclusion

In this research, the hyperparameter tuning of SAEA with CNN is proposed for the prediction and classification of diabetes mellitus. The proposed method is helpful for the detection and classification of diabetes with the knowledge of high-level representation of diabetes pointers. The proposed method is trained by utilizing the three benchmark datasets such as PIDD, Frankfurt Hospital, Germany,

and NCSU datasets for estimating the model's effectiveness. The proposed SAEA with CNN efficiently classifies the diseases into diabetes and non-diabetes. This research compared the effectiveness of the proposed method with the various existing methods. The effectiveness of the proposed method is validated by using performance metrics such as accuracy, precision, recall and F1-score. In comparison results, the proposed SAEA with CNN attains better accuracy results of 99.87% and 99.99% by using PIDD and Frankfurt Hospital, Germany datasets as compared to the previous methods like CNN, Deep 1D-CNN, SAE-CNN, ResNet18 and ResNet50-ReliefF and DCNN with data modelling respectively. In future work, the proposed method will extend to perform different DL approaches for enhancing the overall performance in diabetes prediction.

Notations

Variables	Descriptions
$\hat{x}$	Normalized data
x	Actual data
x_{min} and x_{max}	Minimum and maximum of every feature.
χ^2	Chi-square score
s_{ij}	ith feature value along with instances.
μ_{ij}	Expected count
f	Features
c	Classes
s_{i*}	ith value of particular feature
s_{*j}	number of instances in class j
s	number of instances
$V_{i,G}$	Mutation vector
$X_{i,G}$	Target vector
$U_{i,G}$	Trial vector
$X_{a,G}$, $X_{b,G}$ and $X_{c,G}$	The vectors are arbitrarily chosen from the present population
$u_{i,G}^{j}$, $v_{i,G}^{j}$ and $x_{i,G}^{j}$	jth parameter in ith trial, mutant as well as target vector
$Rand_2$, $Rand_3$, $Rand_4$	independently developed arbitrary numbers in the range between 0 and 1 respectively
$CR_{i,G}$	Crossover mutation
$f(U_{i,G})$	Objective function
$X_{i,G+1}$	vector with minimal objective function survives into further generation
n	Number of layers in the coding network
h	Activation function
m_x and y_k	Input and output feature map
$w_{n,k}$	Kernel weight
$*$	Convolutional operation
b_k	Bias
$r_{i,E}$	Mapping feature
$e_{i:1+z-1}$	Sword embedding vector
$\sigma(.)$	Nonlinear activation function of convolutional operation
°	Hadamard product between two matrices
W	Weight
z	Softmax function input vector contains "n" features of "n" target values
zi	ith item of input vector
e^{z_j}	Standard exponential function
$\sum e^{z_j}$	Normalization term to acquire valid probability distribution
TP	True Positive
TN	True Negative
FP	False positive
FN	False Negative

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] V. Chang, J. Bailey, Q.A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms", *Neural Computing and Applications*, Vol. 35, No. 22, pp. 16157-16173, 2023.

[2] U. Ahmed, G.F. Issa, M.A. Khan, S. Aftab, M.F. Khan, R.A. Said, T.M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning", *IEEE Access*, Vol. 10, pp. 8529-8538, 2022.

[3] M.S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset", *Computer Methods and Programs in Biomedicine Update*, Vol. 4, p. 100118, 2023.

[4] C.C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective", *Computer Methods and Programs in Biomedicine*, Vol. 220, p. 106773, 2022.

[5] R. Rastogi, and M. Bansal, "Diabetes prediction model using data mining techniques", *Measurement: Sensors*, Vol. 25, p. 100605, 2023.

[6] S. Simaiya, R. Kaur, J.K. Sandhu, M. Alsafyani, R. Alroobaea, M. Margala, and P. Chakrabarti, "A novel multistage ensemble approach for prediction and classification of diabetes", *Frontiers in Physiology*, Vol. 13, p. 1085240.

[7] E.K. Oikonomou, and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction", *Cardiovascular Diabetology*, Vol. 22, No. 1, p. 259, 2023.

[8] P. Theerthagiri, A.U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms", *SN Computer Science*, Vol. 4, No. 1, p. 72, 2022.

[9] M.M. Bukhari, B.F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S.S. Ullah, "An improved artificial neural network model for effective diabetes prediction", *Complexity*, Vol. 2021, p. 5525271, 2021.

[10] N. Ahmed, R. Ahammed, M.M. Islam, M.A. Uddin, A. Akhter, M.A. Talukder, and B.K. Paul, "Machine learning based diabetes prediction and development of smart web application", *International Journal of Cognitive Computing in Engineering*, Vol. 2, pp. 229-241, 2021.

[11] K. Abnoosian, R. Farnoosh, and M.H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models", *BMC bioinformatics*, Vol. 24, No. 1, p. 337, 2023.

[12] K.N. Soumya, and R.P. KN, "Diabetes Mellitus Disease Prediction and Classification using Latent Dirichlet Allocation and Artificial Neural Network Classifier", *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 12, No. 10s, pp. 98-106, 2024.

[13] S.A. Alex, N.Z. Jhanjhi, M. Humayun, A.O. Ibrahim, and A.W. Abulfaraj, "Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE", *Electronics*, Vol. 11, No. 17, p. 2737, 2022.

[14] H. Naz, and S. Ahuja, "SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset", *International Journal of Diabetes in Developing Countries*, Vol. 42, No. 2, pp.245-253, 2022.

[15] O.R. Shahin, H.H. Alshammari, A.A. Alzahrani, H. Alkhiri, and A.I. Taloba, "A robust deep neural network framework for the detection of diabetes", *Alexandria Engineering Journal*, Vol. 74, pp. 715-724, 2023.

[16] T. Beghriche, M. Djerioui, Y. Brik, B. Attallah, and S.B. Belhaouari, "An efficient prediction system for diabetes disease based on deep neural network", *Complexity*, Vol. 2021, p. 6053824, 2021.

[17] M.K. Gourisaria, G. Jee, G.M. Harshvardhan, V. Singh, P.K. Singh, and T.C. Workneh, "Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations", *IET Communications*, Vol. 16, No. 5, pp. 532-547, 2022.

[18] M.T. García-Ordás, C. Benavides, J.A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation", *Computer Methods and Programs in Biomedicine*, Vol. 202, p. 105968, 2021.

[19] S.A. Alex, J.J.V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction", *Neural Computing and Applications*, Vol. 34, No. 2, pp. 1319-1327, 2022.

[20] M.F. Aslan, and K. Sabanci, "A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data", *Diagnostics*, Vol. 13, No. 4, p. 796, 2023.

[21] K.K. Patro, J.P. Allam, U. Sanapala, C.K. Marpu, N.A. Samee, M. Alabdulhafith, and P. Plawiak, "An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques", *BMC bioinformatics*, Vol. 24, No. 1, p.372, 2023.

[22] R.V. Aswiga, M. Karpagam, M. Chandralekha, C.S. Kumar, M. Selvi, and S. Deena, "An automatic detection and classification of diabetes mellitus using CNN", *Soft Computing*, Vol. 27, pp. 6869-6875, 2023.

[23] PIMA dataset link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[24] Frankfurt Hospital, Germany dataset link: https://www.kaggle.com/code/linggarmaretva/frankfurt-hospital-diabetes-with-lgbmclassifier.

[25] NCSU dataset link: https://www.kaggle.com/datasets/tmleyncodes/ncsu-diabetes-dataset.

[26] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", *International Journal of Cognitive Computing in Engineering*, Vol. 2, pp. 40-46, 2021.

[27] M.R. Hassan, S. Huda, M.M. Hassan, J. Abawajy, A. Alsanad, and G. Fortino, "Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion", *Information Fusion*, Vol. 77, pp. 70-80, 2022.