DOI: http://dx.doi.org/10.26483/ijarcs.v17i3.7438

Volume 17, No. 3, May-June 2026

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

QUADRATIC MAPREDUCE DOMAIN ADAPTIVE DISCRIMINATIVE AI FOR DIABETES 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
Associate Professor & Head,
PG &Research Department of Computer Science,
Government Arts College (Affiliated to Bharathidasan University, Tiruchirappalli),
Trichy-620022, India.

Abstract **:** Modern healthcare monitoring systems play a vital role in enhancing patients' quality of life by facilitating continuous examination of their health conditions. In healthcare industry, diabetes mellitus is a key risk to public health, and early detection is essential for efficient care and issue prevention. Although numerous deep learning methods have explored smart healthcare systems for diabetes prediction, achieving high accuracy along with time-efficient prediction remains a significant challenge.In order to address these challenges, a novel method called Quadratic MapReduce domain adaptive Discriminative Artificial Intelligence (QMDADAI) is introduced. The main objective of QMDADAI model is to perform accurate diabetes disease prediction with high accuracy and minimal time consumption in healthcare monitoring system. The Discriminative AI model comprises of various processing steps, namely data acquisition, preprocessing, feature selection and classification to enhance the performance of diabetes disease prediction. In the data acquisition phase, the numerous patient data samples are collected from the dataset. Subsequently, the data pre-processing stage is carried out which includes two major processes namely missing data handling and outlier data removal from the input dataset. After data pre-processing, the more important feature selection process is carried out using MapReduce framework from the input dataset. Once the features selected, the classification process is carried out using sequential rank correlation for diabetes disease prediction with higher accuracy. The fine tuning process of Discriminative AI is done using Bats Echolocation optimization algorithm for minimizing the error and enhancing the accuracy of the diabetes disease prediction.Experimental assessmentof QMDADAI model is conducted with different evaluation metrics such as accuracy, precision, recall, F1 score, specificity, ROC-AUC, confusion matrix and training time. The results indicate that the proposed QMDADAI model achieved higher accuracy in diabetes disease prediction with minimal time consumption compared to existing deep learning methods.

Keywords: Diabetes prediction**,** Discriminative AI, Map Reduce framework, sequential rank correlation, Bats Echolocation optimization algorithm

1. INTRODUCTION

Diabetes is a metabolic state that leads to cause a chronic illness and organ malfunction if it remains not treated at an earlier stage.Diabetes is a medical condition characterized by eminent blood glucose levels away from the normal range. Accurate detection of this disease is vital in the healthcare sector for efficient treatment and supervision.In general, this Diabetes disorder obvious itself as high blood glucose levels caused by either inadequate insulin construction (Type 1 diabetes mellitus (T1DM)) or alack of abilityto efficiently uses insulin (Type 2 diabetes mellitus (T2DM)). Several deep learning methods are more helpful for diagnosis of diabetes and analyze complex medical data, uncover subtle patterns and enhance risk assessment.

Serial Cascaded Convolutional Ensemble Network (SCCEN) was developed in [1] with the aim of performing diabetes detection at an earlier based on optimal weighted feature. The developed model failed to develop more accurate and comprehensive diagnostic models by adding robust data preprocessing techniques to improve the model's performance on diabetes detection. A hybrid Convolutional Neural Network-Bidirectional Gated Recurrent Unit (CNN-BiGRU) model was introduced in [2] for Diabetes Prediction using Adaptive Synthetic Sampling (ADASYN) based data augmentation. However, a scalable and reliable methodology was not implemented for diabetes diagnostics aimed at improving predictive capabilities in other medical domains.

Secretary Bird Optimization Algorithm (QHSBOA) was developed in [3] by integrating Kernel Extreme Learning Machine (KELM) for diabetes classification. However, the algorithm failed to enhance the accuracy and robustness of problem-solving.A novel machine learning classifiers were developed in [4] for accurate diabetes prediction precisely from limited labeled data and outliers in diabetes datasets. But the model did not train using several deep learning methods to provide improvements on diabetic patients.Machine learning (ML)-based techniques were developed in [5] for early detection of Gestational Diabetes Mellitus. However, the technique was relatively time-consuming and expensive process.

Multi-class classification framework was developed in [6] to classify patients across type 2 Diabetes Mellitus (T2DM) without complications. However, validation on larger, multi-center datasets with diverse was not considered. Recursive Feature Elimination-GRU (RFE-GRU) was developed in [7] based on selecting the relevant features. However machine learning and deep learning models was not applied to acquire better results. A new two-stage ensemble framework combining Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) were developed in [8] with randomized feature selection to increase the diabetes prediction accuracy. However, computational scalability of the model was not improved.

An ensemble univariate short-term predictive model was introduced in [9] for type 1 diabetes and minimizing its error. However, physiological parameters from medical devices were not considered for type 1 diabetes prediction. A new Sensor-infused Quantum CNN was introduced in [10] for diabetes Identification and Diet recommendation at early stage using an IoT system. However,computational approaches were not employed by the model to enhance its ability to forecast diabetes level more precisely and effectively.

An adaptive ensemble feature selection technique was introduced in [11] for model-agnostic diabetes prediction. However, it failed to explore methods for feature interpretation, visualization, and domain-specific validation to enhance performance and stability.Machine Learning model was developed in [12] for Diabetes Risk Prediction based on supervised filter for feature selection. However, it failed to extend interpretability and computational efficiency.Random Forest and Extreme Gradient Boosting (XGBoost) classification algorithms were developed in [13] aimed to improve accurate diabetes prediction accuracy through feature selection techniques. However, methods were not efficient, especially for larger, more diverse datasets. Explainable AI-IoT based diabetes prediction was employed in [14] by means of TabTransformer model. However, it failed to contribute towards AI driven diabetes monitoring system to ensure accuracy for real world healthcare applications. A personalized T2DM risk prediction model was introduced in [15] using machine learning (ML) with the aim of type 2 diabetes risk prediction. However, it failed to predict long-term outcomes such as complications and mortality.

1.1 Major Contributions

The key contributions of the QMDADAI model are outlined as follows:

- To design a newmethod calledQMDADAI modelfor accurate diabetes disease prediction by integrating four various processes namely data pre-processing, feature selection, classification and fine tuning.
- To reduce the accurate diabetes disease classification time, QMDADAI model is introduced which integratesdata pre-processing and significant feature selection based on MapReduce framework. This process reduces the time consumption of diabetes disease classification time.
- To enhance the accuracy of diabetes disease classification, a sequential rank correlation method is employed in Discriminative AI model to analyze data samples. Moreover, the adaptive Bats Echolocation optimization algorithm is applied for hyperparameter fine-tuning to further reduces the error rate and increases the precision as well as recall during diabetes disease classification.
- The relative analysis of the proposed QMDADAI model has been evaluated based on conventional deep learning methods with dissimilar performance metrics.

1.2 Structure of paper:

The rest of the paper is organized as follows. Section 2 presents a detailed survey of previous studies on diabetes disease prediction. Section 3 introduces the proposed a QMDADAI model, describing its architecture and overall workflow with the aid of a diagram. Section 4 outlines the experimental setup, including a description of the dataset used. Section 5 provides a thorough comparative evaluation of the proposed QMDADAI modelagainst existing methods using various performance measures. Finally, Section 6 summarizes the main findings and concludes the study.

2. Related works

Stacking machine learning optimized by Grey Wolf Optimizer was developed in [16] to construct and evaluate prediction models for type 1 diabetes patients.However, timely diabetes self-management was not employed to decrease complications. An advanced imputation strategies and ensemble machine learning model were developed in [17] to enhance the diabetes disease prediction accuracy. However, it failed to extend the study by validating the proposed ensembleframework on larger and more diverse datasets.Grey Wolf Optimizer with an integrated Autophagy Mechanism (GWO-AM) was developed in [18] to improve prediction accuracy while minimizing the number of selected features. However, it did not integrating with deep learning models to enhance prediction accuracy. Light Gradient Boosting Machine (LightGBM) was introduced in [19] aimed to enhance the predictive accuracy of diabetes. However, it failed to integrate diverse data types to achieve improved accuracy and robustness in predicting diabetes risk.Transformer encoder model was designed in [20] for detecting the diabetes disease diagnosis by analyzing longitudinal heterogeneous diabetes data. However, model was not suitable for larger datasets to enhance the fine-tuning performance.

Machine learning-based predictive model was developed in [21] using XGBoost algorithm for diabetes disease detection to improve model performance. However, it failed to investigate the incorporation of this recall optimized to improve the efficiency. Random forest machine learning algorithms were designed in [22] for precise measurement and risk stratification of Type 1 diabetes. However, timely screening and monitoring risk and disease progression from an early stage.An optimized Light Gradient-Boosting Machine (Light GBM) algorithm was developed in [23] for predicting diabetic progression of Type 2 Diabetes. However, the designed algorithm failed to diabetic progression. An advanced machine learning techniques were introduced in [24] for enhancing the diabetes prediction. However, computationally intensive and time-consuming, especially with complex models was not reduced.

Machine learning techniques were designed in [25] for early detection and management of diabetes based on significant features. However, advanced and deep learning algorithms were not developed for more robust and reliable diabetes prediction. An ensemble deep learning model was developed in [26] for the diagnosis of diabetes to improve the average accuracy and precision. However, it failed to enhance the quality of disease diagnosis. Associative Kruskal Wallis and MapReduce Poly Kernel (AKW-MRPK) method was introduced in [27] for early disease prediction. However efficient optimization techniques were not implemented to extract the significant features from the large dataset for providing better results. An optimized XGBoost model with Bayesian optimization was designed in [28] to improve outcomes for individuals at risk of developing diabetes.However approach failed to diabetes prediction with high accuracy. Parallel and sequential

ensemble machine learning approaches were developed in [29]to boost diabetes classification accuracy based on feature selection. However, it failed to improve prediction accuracy of different types of diabetes.XGBoost and explainable AI model were introduced in [30] for accurate prediction of type 2 diabetes. However, it failed to focus on incorporating deep learning architectures for minimizing the prediction error.

3. Proposal methodology

Diabetes mellitus is a widespread chronic sickness with severe complications that require timely diagnosis. Accurate and timely detection of diabetes disease and efficient treatment are essential to decrease the effects of diabetes. This paper introduces a QMDADAI model for early diabetes diseasesdetection. A novel approach transfer learning called Discriminative AIis introduced for precise detection of diabetes disease. The proposed Discriminative AI is to enhance accuracy of diabetes disease.

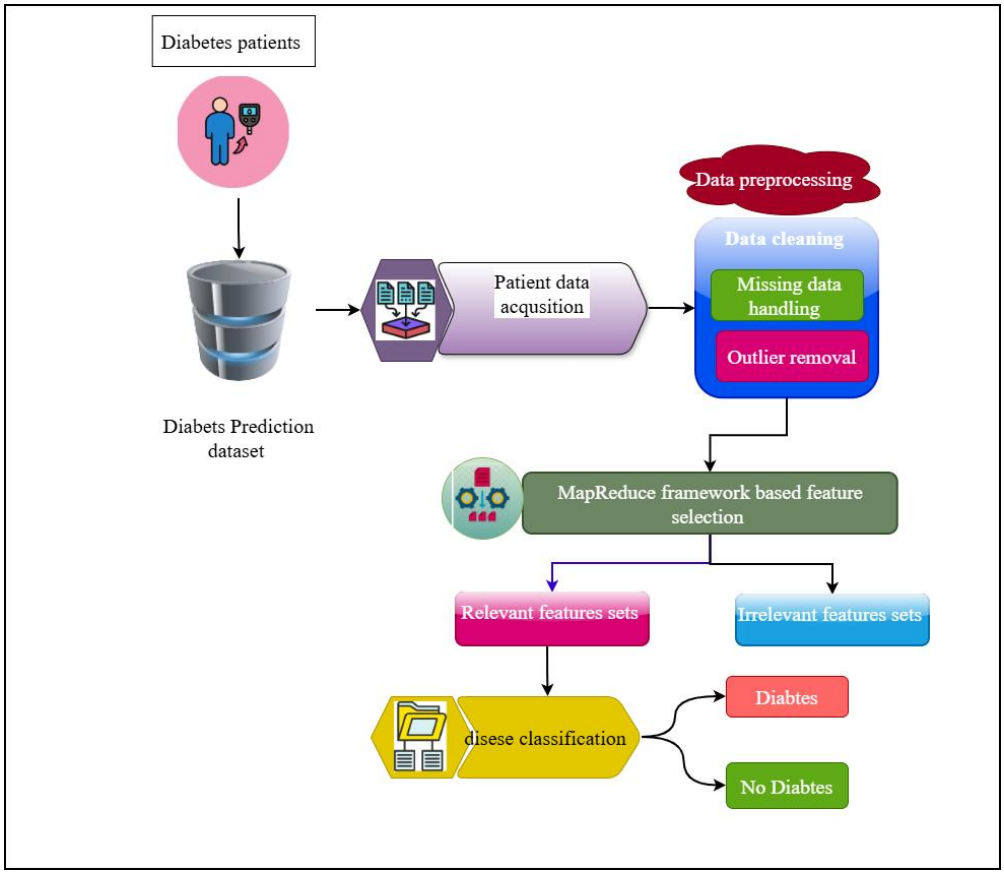


Figure 1 overall structural design of proposed QMDADAI model

Figure 1 depicts the overall processing diagram of proposed QMDADAI model for accurate diabetes disease prediction. In order to achieve the accurate diabetes disease prediction from the dataset, the proposed model consists of different processing steps namely Data acquisition, pre-processing, feature selection, and classification into the construction of Discriminative artificial intelligence network. To start with, proposed QMDADAI model collects the numerous patient data samples from the kaggle repository dataset for accurate diabetes prediction. Subsequently, proposed model performs the data pre-processing or cleaning for converting the raw diabetes prediction dataset into more suitable format by the means ofhandling missing data and outlier data removal. Afterwards, significant features selection process is carried out by employing Mapreduce framework for accurate diabetes classification with minimal time consumption. To end with, the proposed QMDADAI model executes the diabetes prediction through binary classification with the help of sequential rank correlation attaining higher accuracy with lesser prediction error. The succeeding sections briefly represent the fundamental processes of the proposed QMDADAI model, highlighting the principle and operational implication of each step.

3.1 Data acquisition

The primary step of proposed QMDADAI model is to perform the data collection which involves gathering the raw patient data samples from the Diabetes prediction dataset https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset.The dataset comprises of patient's medical and demographic data in addition to their diabetes status (positive or negative). This dataset includes 9 features (age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level and target variable and 100000 data samples. The feature descriptions are listed in table 1.

Table 1 Attribute description

S.No	Attributes	Description
1	Gender	Gender of patient Female, Male

2	Age	Age of patient
3	Hypertension	Indicates whether a patient has hypertension
4	Heart Disease	Whether a patient's has heart diseases (0:yes, 1:no)
5	Smoking history	It describes an individual's current or past Smoking behavior
6	Body Mass Index (BMI)	It is a numerical value calculated from a person's weight and height
7	HbA1c_level	It indicates whether a person's average blood sugar level over the past 2-3 months
8	Blood glucose level	It refers to the mean concentration of glucose present in the blood
9	diabetes	Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0

3.2 Domain adapted discriminative artificial intelligence network

A Discriminative Artificial Intelligence (AI) is a types of deep learning model designed to categorize input data into predefined categories by utilizing the direct relationship between input features and output labels. In this context, transfer learning is a Domain adapted discriminative artificial intelligence network that utilizes the previously learned knowledge from one task to enhance the learning process. In contrast to traditional deep learning approaches, transfer learning facilitates reduced training time and often achieves improved performance, especially when handling large-scale data samples, thereby enhancing overall efficiency.

Domain adaptation addresses the challenge of training a model on one data distribution in the source domain and applying it to a related but different data distribution in the target domain. The main advantage of domain adaptive Discriminative AI reduces computational cost and data collection effort for large-scale applications. The structure of Domain adapted discriminative artificial intelligence model is shown in figure 2.

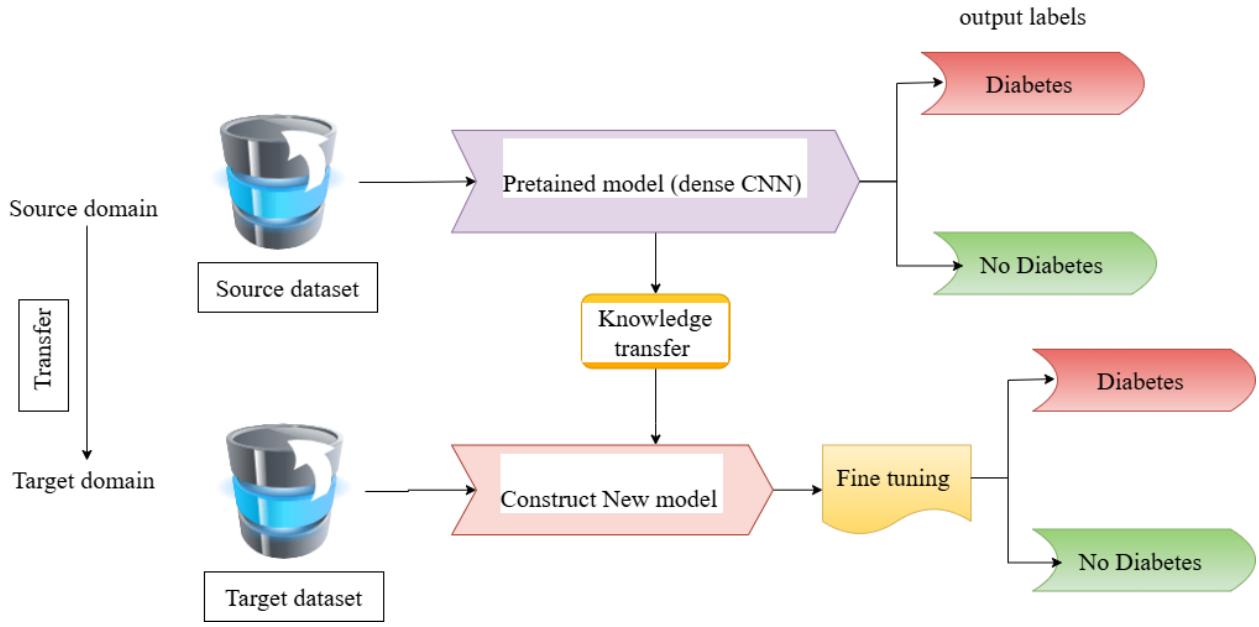


Figure 2 Schematic structure of Domain adapted discriminative AI network

Figure 2 illustrates the graphic structure of a Domain adapted discriminative AI network, which consists of two types of domain such as source and target. The source domain utilizes the pre-trained model which utilizes the denseCNN. The target domain utilizes the new model which utilizes the same denseCNN but the data distribution is different. In addition, create a new model utilizes the knowledge transfer from the existing pre-trained model. However new task-specific fine-tuning is performed to generate the binary classification results, as illustrated in figure 3.

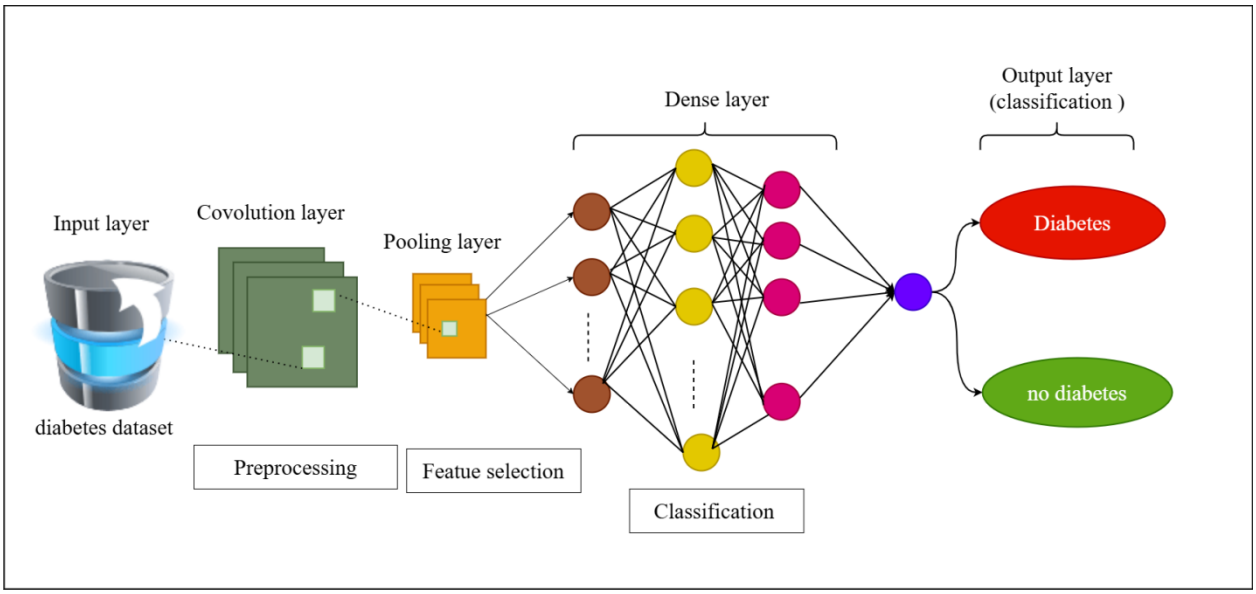


Figure 3 Structure of DenseCNN

Figure 3 demonstrates the structure of a DenseCNN to perform the classification of diabetes diseases. As shown in figure 3, the input and output layers each consist of a single layer, whereas the hidden layers are composed of three sub-layers, including convolutional (Conv) layers, max-pooling layers, and dense layers. Neural networks are made up of small computational units called artificial neurons or nodes for receiving the input signals, process them, and transferred the resulting output to neurons in the next subsequent layer.

Let us consider the input data matrix is given as input to the input layer of deep learning architecture. It is formulated as given below,

$$z = \begin{bmatrix} F_1 & F_2 & \dots & F_m \\ D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \vdots & \vdots & \dots & \vdots \\ D_{m1} & D_{m2} & \dots & D_{mn} \end{bmatrix} \qquad m = rows, n = column \tag{1}$$

Where,z represents an input data matrix where each column represents a number of features$F = \{F_1, F_2, \dots, F_m\}$, each row represents a number of data samples or records '$D = \{D_1, D_2, \dots, D_n\}$' respectively. These input data is sent into the first hidden layer of where the neuron activation probability is measured. Therefore, the neuron activation probability in the hidden layer 'P_{Conv}' is expressed as follows,

$$P_{Conv} = \sum(z * \beta_{ih}) + b \tag{2}$$

Where,P_{Conv} denotes a neuron activation probability in the convolution layer, z represents a input matrix associated to a weights between input and hidden layer 'β_{ih}' and with bias 'b' helps to transfer the input data samples

Following this, a data pre-processing step is performed in convolutional layer to address missing values and remove outliers present in the dataset. During this stage, the raw input data is transformed into a more suitable and structured format to improve the accuracy of diabetes disease prediction. In pre-processing step, two major processes namely missing data imputation and outlier removal are carried out to organize the dataset.

. In this imputation method, polynomial linear regression imputation is applied to determine the missing value within the dataset based on the available data points. Therefore, the regression process is expressed as follows,

$$D_M = \vartheta_0 + \vartheta_1 D_1 + \vartheta_2 {D_2}^2 + \vartheta_3 {D_3}^2 \dots \vartheta_n {D_n}^n \tag{3}$$

Where, D_M represents a determined missing data value, $\vartheta_0, \vartheta_1, \vartheta_2 \dots \vartheta_n$ represents a polynomial coefficients with available data points $D_1, D_2, \dots, D_n$. This identified missing data values 'D_M ' are imputed into the null entries within the dataset.

After that, the outlier's data are identified within the dataset to enhance the accuracy of the deep learning model. Outlier data refers to data points that considerably diverge from the rest of the data samples within the dataset. The proposed method utilizes the robust mean absolute divergence model is developed to detect the outlier data samples.

First, the given input data samples are organized in an ascending order. Subsequently, the median value 'D_{Med}' is calculated. Then the mean absolute divergence is calculated with the median values.

$$MAD = \omega\left(\frac{|D_i - D_{Med}|}{\sigma_{MA}}\right) \tag{4}$$

Where, MAD indicates a mean absolute divergence,$D_i$indicates a data, D_{Med} represents a median value of samples, σ_{MA} indicates a mean absolute deviation, $|D_i - D_{Med}|$ represents a absolute divergence between the data and their median value, ωindicates a scaling factor and the value is 0.6745. The outcomes ofMAD returned from 0 to 1.

$$K = \begin{cases} MAD > T\,, & Outlier\ data\ samples \\ MAD \le T\,, & not\ outlier \end{cases} \tag{5}$$

Where, divergence 'MAD' value is larger than the threshold 'T', the data sample is said to be an outlier. Otherwise, data is said to be an outlier. In this preprocessing step, all outlier data points present in the diabetes dataset are detected and removed to improve the accuracy of predictions. The cleaned and preprocessed data is then transferred to the pooling layer for selecting the more pertinent features from the dataset.

- **Pooling layer (feature selection)**

The second process of pre-trained model is the feature selection to reduce the overall dimensionality of the

dataset for accurate diabetes disease detection with minimal time consumption. The proposed model utilizes the MapReduce framework is a distributed programming model designed to process and analyzes large-scale datasets effectively by partitioning the complex tasks into smaller sub-tasks and executing them in parallel.

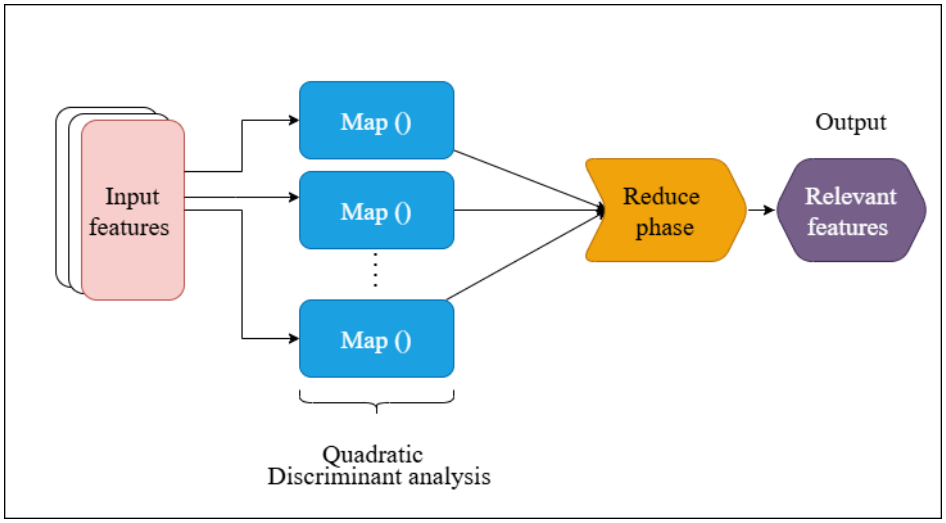


Figure 4 MapReduce framework

Figure 4 illustrates the MapReduce framework designed to process and analyzes large-scale datasetsand executing them in parallel.MapReduce consists of two primary phases namely Map phase and the Reduce phase. In the Map phase, input data i.e. number of features is processed in parallel. In the Reduce phase, the relevantfeatures are aggregated to produce the final outputi.e. dimensionality reduced output.

During the map phase, featured are analyzed using the Quadratic Discriminant analysis based on various features. The mapper processes input features and converts it into key and value pairs.

$$Map\left(F_j\right) \rightarrow (F_j, Q) \quad (6)$$

Where, $F_j$input features i.e. key value, Q represents the value (i.e. similarity score). The similarity score between the features and target is identified by applying Quadratic Discriminant analysis. The Quadratic Discriminant analysis is a dimensionality reduction methodhelps to analyze the likelihood between the features and target values formulated as follows

$$Q = \left[exp\left(-0.5 * \sum_{j=1}^{m}\left(\frac{|F_j - F_T|^2}{var^2}\right)\right)\right] \quad (7)$$

Where, Q indicates an outcome of Quadratic Discriminant analysis, $arg\ max$ denotes arguments of maximum probability output between the feature 'F_j' and target features (F_T), 'var 'specifies a standard deviation between the features and target.

The feature analysis process identifies the most significant features by evaluating their maximum outcome values. Features with minimal probabilistic significance are eliminated from the dataset. Finally, the significant features are retained at the reduce phase.

$$Reduce\left(F_j\right) \rightarrow \arg\max(Q) \quad (8)$$

In the Reduce phase of the MapReduce framework, the Reduce function then applies an $\arg max$ on Q to identify and select the features that exhibit the highest likelihood. Thus, the Reduce phase ensures that only the most relevant features are forwarded for subsequent classification tasks.

- **Dense layer (classification)**

After selecting the significant features from the dataset, the final step of proposed is to perform the classification in dense layer of network architecture together with the selected features. In the dense layer, training and testing data analysis is done by the means of Sequential rank correlation to classify the diabetes diseases for accurate recommendation.

$$SRC = 1 - 6 * \frac{\sum Diff^2}{(n(n^2-1))} \quad (9)$$

$$WSS = \frac{|D_{Tr} - D_{Ts}|^2}{|D_{Tr} - D_{Ts}|^2 + \rho^2} \quad (10)$$

$$\rho = \frac{D_{Tr} - D_{Ts}}{\sqrt{(1-D_{Tr})^2(1-D_{Ts})^2}} \quad (11)$$

Where, SRCdenotes a Sequential rank correlation, D_{Tr} denotes a training data samples, D_{Ts} denotes a testing data samples, $Diff$ indicates difference between the rank. The rank is assigned based onwald statistic test statistics 'WSS'. The output of WSS' is close to high score and assign the high rank. Otherwise, assign the low rank. The correlation 'SRC'returns the output ranges from '0 to '1'. If the outcome of 'SRC' is more close to 1, indicates that the disease diagnosis have been accurately predicted.

Finally, the disease prediction results are obtained at the output layer as follows,

$$Y = F_{sig}\left(\beta_{ho} * h_t\right) \quad (12)$$

Where Y indicates a binary classification output, F_{sig} indicates a sigmoid activation function, h_t indicates an output of the previous hidden layer, β_{ho} denotes a weight between the hidden and output layer. Sigmoid activation function 'F_{sig}' in the output layer is employed for binary classification output such as diabetes and no diabetes. The algorithmic process of disease diagnosis using pre-trained model is given below.

// Algorithm 1: Pre-trained classification model

Input: **Dataset, features** $F_1, F_2, \dots, F_m$ **and data samples**$\{D_1, D_2, \dots, D_n\}$
Output: classification of diabetes diseases

Begin
Construct the pre-trained model

1. **Collect n**umber of data samples $D_1, D_2, \dots, D_n$ **at the input layer**
2. **For each training** samples D_i **--- convolutional layer**
3. Compute the neuron activation probability using (2)
4. Handling missing data according to (3)
5. **For each data** samples D_i
6. **Compute** themean absolute divergence using (4)
7. **If** $(MAD > T)T$**then**
8. return outlier data point
9. **else**
10. return outlier data point
11. **End if**
12. **End for**
13. **End for**
14. **For each pre-processed dataset --- Pooling layer**
15. Apply Quadratic Discriminant analysisbetween features using (9) in map phase
16. **If** $(\boldsymbol{Max}\ Q)$ **then**
17. Indicates a high degree of similarity
18. Select the relevant features
19. **else**
20. Remove irrelevant features
21. **End if**
22. **Return the** relevant features in reduce phase using (8)
23. **End for**
24. **For each** data samples **do--- --- Dense layer**
25. Measure the Sequential rank correlation using (9)
26. Obtain final results using sigmoid activation function **--- output layer**
27. **If**$Y = 1$**then**
28. Data sample classified as diabetes
29. **Else**
30. Data sample classified as normal
31. **End if**
32. **End for**

End

Algorithm 1 describes the process used for diabetes disease detection through discriminative AI model. A large volume of patient data samples is gathered from the dataset for training process. A dense CNN used as the pre-trained model consists of numerous layers namely convolutional, pooling layer and dense layer to process and learn from the input data samples. To begin with, the training samples are given to the input layer. The weighted sum is calculated in the following convolutional layers. The first convolutional layer performs data pre-processing which executes missing data handling and outlier's removal. Afterward, relevant selection process is carried out using the Mapreduce frameworkto preserve only the more relevant features and remove the others in reduce phase. The correlation function is employed in third hidden layer to analyze the training and testing data samples. Finally, the output layer utilizes a sigmoid activation function to produce the final diabetes disease diagnosis outcomes.

3.2.2 Construct the new classification model for accurate diabetes disease diagnosis

In a discriminative AI model, a collection of input training samples is used to create a new classification model with different data distributions. In order to build a new model, already trained networkcalled Dense CNN. As illustrated in Figure 2, the frozen layers symbolize the pre-trained section of the network, where the initial two layers are kept fixed and are not updated during training. These layers preserve the knowledge collected from the base model and serve as a feature extractor. To adapt the model to a new task, fine-tuning process of dense layer is introduced. The input data is first propagated through the frozen layers to perform feat processing and extract relevant features for diabetes disease diagnosis. By keeping these layers unchanged, the model achieves faster training and minimizes the computational complexity, thereby improving overall efficiency.

To construct a new model, the discriminative AI model architecture retains the input layer and the first two hidden layers by freezing them and it added for fine-tuning and output. First, set of input training samples is given to the input layer of Dense CNN and it transferred to hidden layers, where initial data pre-processing is completed. This includes handling missing values using polynomial linear regression imputation using (3) and detecting outliers with the mean absolute divergence using (4) method. In the second hidden layer, relevant features are selected using the

Map reduce function using (6) (7) (8). Finally, disease detection is performed in the dense layer by applying the sequential rank correlation to classify the diabetes diseases using (9).

The diabetes diseases output from the hidden layer is fine tuned by applying Accelerated Gradient Descent to adjust the weights of corresponding layers. Finally to improve the learning rate of the network, Bats Echolocation optimization is utilized in the hidden layer that in turn measures the derivatives of weights. It was inspired by the echolocation behavior of microbats for finding tan optimal weight thereby reducing the disease diagnosis errors and improving the accuracy. For each classificationresult, theerror rate is measured as follows,

$$\varepsilon = \left(Y_{act} - Y_{predicted}\right)^2 \quad (13)$$

Where, Y_{act}indicates an actual output, $Y_{predicted}$denotes the predicted output, ε denotes a disease diagnosis errors. In the context of fine-tuning phase, the hyperparameters of dense layer of network model are adjusted by means of Accelerated Gradient Descent algorithm. This algorithm helps to iteratively adjust the weights for quicker and more efficient convergence. The initial hyperparameters ' β'gets updated by using adaptive momentum rule as follows

$$\beta_{t+1} = \beta_t - \eta\left(\frac{\partial\varepsilon}{\partial\beta}\right) \quad (14)$$

Where, β_{t+1} denotes an updated hyper parameter, β_t indicates a current hyperparameter, η denotes a learning rate, and $\frac{\partial\varepsilon}{\partial\beta}$' representsthe partial derivative of error rate '$\varepsilon$with respect to the weight 'hyperparameters, β'. This iterative procedure continues until the prediction error reaches its minimum value. To achieve this, the optimal set of hyperparameters is determined using a Bats Echolocation optimization algorithm. This metaheuristic technique is capable of exploring the global search space effectively by mimicking the natural hunting behavior of bats.

The proposed Bats Echolocation optimization algorithm initializes the populations of the bats (i.e hyperparameters) are moved around in the search with the initial position $P_i(t)$and velocity$H_i(t)$.

$$\beta = \beta_1, \beta_2, \ldots . \beta_r \quad (15)$$

Where, β indicates a weights between the layers of dense layer of deep learning as architecture. After initializing the population, the fitness is computed along with the following expression,

$$fitness\ (\beta) = \arg\min\ \varepsilon \quad (16)$$

Where, $fitness\ (\beta)$ indicates the fitness function, $\arg min$ denotes an argument of the minimum function, 'ε' represents an error. If the fitness of bat is greater than others, then the optimal one is selected by applying a stochastic universal sampling selection process. The selection of individuals is carried out based on probability assessment as follows,

$$P = \frac{fitness\ (\beta_l)}{\sum_{l=1}^{b} \beta_l} (17)$$

Where, selection probability (P) is evaluated based on the ratio of every individual fitness $'fitness\ \beta_l$ ' to the average fitness of the population in j^{th} individual 'φ_{F_j}'.It means the best individuals are chosen with high probability 'P'.

$$P_i(t+1) = P_i(t) + H_i(t+1) \quad (18)$$

$$H_i(t+1) = H_i(t) + 0.5 * |P_i(t) - P_{best}| * K \quad (19)$$

$$K = F_{min} + (F_{max} - F_{min})\,.r \quad (20)$$

Where, $P_i(t+1)$ denotes an updated position of the bat, $P_i(t)$indicates a current position,$H_i(t+1)$ indicates an updated velocity of the bat, $H_i(t)$ indicates the current velocity of the bat,P_{best} indicates a global best solution, r denotes a random vector with uniform distribution, $F_{min} = 0, F_{max} = 1$.

Consequently, optimally selected weights are used to enhance the disease prediction. Finally, the disease prediction results are obtained at the output layer as follows,

$$Y_{new} = F_{sig}\ (\beta_{ho} * h_t) \quad (21)$$

Where, Y_{new} designates a binary classification output, F_{sig} specifies a sigmoid activation function, h_t indicates an output of the previous hidden layer, β_{ho} indicates a weight between the hidden and output layer. The algorithmic process of new classifier model is given below.

// Algorithm 1: new classifier model

Input: dataset 'DS' , number of features$F = \{F_1, F_2, \ldots, F_n\}$, samples '$S = \{S_1, S_2, \ldots, S_n\}$'
Output: Increase the disease prediction accuracy

Begin

1. **Collect n**umber of features$F = \{F_1, F_2, \ldots, F_n\}$ and samples '$S = \{S_1, S_2, \ldots, S_N\}$--- **input layer**
2. **For each** sample 'S'
3. Formulate the neuron activation probability using (1)
4. **End For**
5. processing **the data samples using (3) (4)—[convolutional layer]**
6. **For each pre-processing samples**
7. Perform relevant feature selection using (6) (7) (8)
8. **End for**
9. **For each** training samples and testing samples
10. Measure the sequential rank correlation using (9)
11. **End for**
12. **For each** classification results
13. Measure the error rate 'ε'using (13)
14. **Apply** accelerated gradient descent to update weight using (12)
15. **End for**
16. Initialize the population of the weights $\beta = \beta_1, \beta_2, \ldots . \beta_r$
17. **foreach** weight in populations

18. Compute the fitness '$fitness\ (\beta)$' using (16)
19. While ($t<t_{max}$)
20. Select the current best based on high probability using (17)
21. Generate new location using (18)(19) (20)
22. $t = t + 1$
23. go to step 19
24. E**nd while**
25. Find the optimal weight
26. Obtain the final classification results using (21) with sigmoid activation function **at output layer**
End

Algorithm 2 outlines the process for classifyingdiabetes diseases with minimal time consumption. For each input data sample, weights and biases are assigned to the input layer of the deep learning architecture. The input is then transferred to the neurons of the convolutional layer, where data preprocessing is carried out to handle missing values and outlier detection within the dataset. Subsequently, significant features are selected in the consequentpooling layer. Finally, classification is performed in dense layer using the sequential rank correlation to compare training and testing data samples. Based on this correlation measure, different diabetes diseases are classified. Once the disease classification is performed, a fine-tuning process is employed using the Bats Echolocation optimization. Initially, the number of weights is initialized, and a population of bats (representing the weights) is initialized within the search space. The fitness of each bat is evaluated using the classification error as the objective measure. Based on this evaluation, the position of each bat is updated iteratively to explore better solutions within the search space. This updating process continues until a predefined maximum number of iterations are reached. Through this repeated optimization process, the bats algorithm effectively determines the optimal weight parameters that reduce the classification error. Finally, optimal classifications are obtained by selecting the model configuration that achieves the lowest classification error at the output layer.

4. Experimental Settings

In this section, experimental assessment of the proposed QMDADAI model with existing approaches, SCCEN [1] and CNN-BiGRU [2] are implemented using Java language. In order to conduct the experiment, Diabetes Prediction dataset taken 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. This dataset includes necessary information for accurate diabetic detection and it also including the features about. The dataset includes 100000records and 9 features or attributes. The more description about the feature descriptions are listed in table1.

4.1 Implementation results

The QMDADAI model is widelyanalyzed to calculate its efficiency in diabetes disease prediction. The evaluation comprises many stages, including data acquisition, preprocessing, feature selection, and classification, utilizing the diabetes disease prediction dataset taken from Kaggle repository. First, the 100000 data samples aregathered from the dataset for evaluating the performance of QMDADAI model.

After data acquisition, preprocessing stage is employed, including handling missing values and removing outliers. Initially, the total dataset size is 3721.05KB, which was reduced to 3476.96 after preprocessing. The preprocessing results are shown in figure 5 and 6.

```
Input File Size: 3721.05 KB
Header: gender,age,hypertension,heart_disease,smoking_history,bmi,HbA1c_level,blood_glucose_level,diabetes
Loaded Records: 100000

--- BMI Imputation ---

--- Outlier Detection (Glucose) ---
Outlier Removed: 260.0
Outlier Removed: 300.0
Outlier Removed: 280.0
Outlier Removed: 280.0
Outlier Removed: 300.0
Outlier Removed: 280.0
Outlier Removed: 300.0
Outlier Removed: 280.0
Outlier Removed: 260.0
Outlier Removed: 280.0
Outlier Removed: 260.0
Outlier Removed: 260.0
Outlier Removed: 300.0
Outlier Removed: 240.0
Outlier Removed: 240.0
Outlier Removed: 300.0
Outlier Removed: 280.0
Outlier Removed: 280.0
Outlier Removed: 300.0
Outlier Removed: 300.0
Outlier Removed: 300.0
Outlier Removed: 280.0
Outlier Removed: 260.0
Outlier Removed: 240.0
```

Figure 5 before data preprocessing

```
Remaining after cleaning: 97326

Preprocessed data saved to: processed_diabetes.csv
Output File Size: 3476.96 KB

--- Sample Processed Data ---
0       80.0    0.0     1.0     0       25.19   6.6     140.0   0.0
0       54.0    0.0     0.0     4       27.32   6.6     80.0    0.0
1       28.0    0.0     0.0     0       27.32   5.7     158.0   0.0
0       36.0    0.0     0.0     1       23.45   5.0     155.0   0.0
1       76.0    1.0     1.0     1       20.14   4.8     155.0   0.0
0       20.0    0.0     0.0     0       27.32   6.6     85.0    0.0
0       44.0    0.0     0.0     0       19.31   6.5     200.0   1.0
0       79.0    0.0     0.0     4       23.86   5.7     85.0    0.0
1       42.0    0.0     0.0     0       33.64   4.8     145.0   0.0
0       32.0    0.0     0.0     0       27.32   5.0     100.0   0.0
```

Figure 6 After data preprocessing

After preprocessing, the QMDADAI model focuses on selecting more relevant features that enhance the accuracy of diabetes disease prediction. To attain this, it combinatorial MapReduce framework with Quadratic Discriminant analysis is employed to select 5significant featuresfrom 9 features.

```
--- Feature Importance ---
HbA1c -> 1.135218011302163
age -> 0.6059828492351452
blood_glucose -> 0.33482199436631604
bmi -> 0.26360726702618426
hypertension -> 0.14766484058421275
heart_disease -> 0.09416105828715747
smoking_history -> 0.04015639051807616
gender -> 0.009923129960527055

--- Selected Features ---
HbA1c
age
blood_glucose
bmi
hypertension
```

Figure 7 selected 5 features

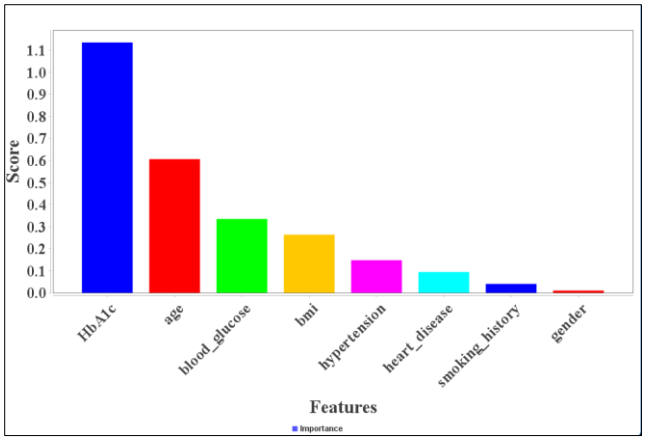


Figure 8 feature importance chart

The figure 8 illustrates the importance scores of different features used in the diabetes prediction model. Among all the attributes, HbA1c has the highest importance score, indicating that it plays a dominant role in identifying diabetes. This is followed by age and blood glucose, which also contribute considerably to the prediction process. **F**eatures such as BMI, hypertension, and heart disease illustrate moderate influence. Meanwhile, smoking history and gender have relatively low significance scores, implying a minimal impact on the model's predictive capability.

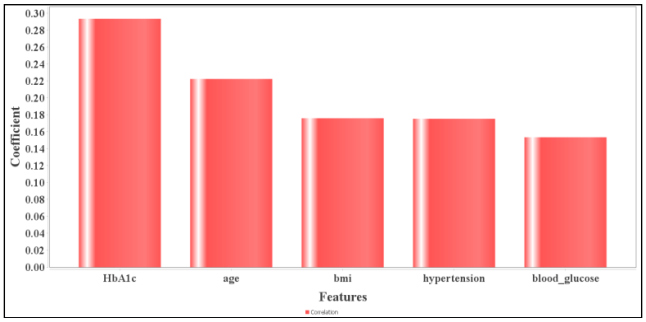


Figure 9 selected feature chart

Figure 9 represents a feature correlation bar chart that illustrates the relationship between selected features and the target variable in the diabetes prediction model. The x-axis displays the features, including HbA1c, age, BMI, hypertension, and blood glucose, while the y-axis shows their corresponding correlation coefficients. Each bar indicates the strength of association between a feature and the outcome. Overall, the graph highlights the relative importance of these features in contributing to the diabetes prediction task.

```
Training Completed

--- Predictions ---
Actual: 0.0 Predicted: 0.12423316505714992
Actual: 0.0 Predicted: 0.008799635219850313
Actual: 0.0 Predicted: 0.0045004536489564395
Actual: 0.0 Predicted: 0.002645877923899969
Actual: 0.0 Predicted: 0.023098020947940338
Actual: 0.0 Predicted: 7.899228733732298E-4
Actual: 1.0 Predicted: 0.012834435311689258
Actual: 0.0 Predicted: 0.024156095651100665
Actual: 0.0 Predicted: 0.012862122026522909
Actual: 0.0 Predicted: 0.0011863538219039685
```

Figure 10 diabetes classification

Figure 10 shows the prediction output of a diabetes classification model, where 0 indicates no diabetes (non-diseased) and 1 represents the presence of diabetes (diseased). Each row compares the actual class label with the predicted probability generated by the model.Overall, the predictions suggest that the model performs effectively in classifying both non-disease and disease cases.

5. **Performance comparative analysis**

In this section, the performance of the proposed QMDADAI model and two existing methods namely SCCEN [1] and CNN-BiGRU [2] are analyzed using various metrics, such as accuracy, precision, recall, F1 score, specificity, ROC-AUC, confusion matrix and training time across different data samples are discussed with the help of table and graphical representation.

Accuracy: Accuracy in diabetes disease prediction refers to the measure of correctly a model based on input data samples. Mathematically, it is measured as the ratio of the number of correct disease classification (true positives and true negatives) to the total number of predictions (including false positives and false negatives).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (22)$$

Where, accuracy 'Acc' is measured based on the true positive rate (TP), which reflects the number of diabetes diseased samples correctly classified, false positive rate (FP), representing no- diabetes samples incorrectly classified as diseased, false negative rate (FN), indicating diseased samples incorrectly classified as disease, and the true negative rate (TN), which denotes the correct detection of diabetes disease prediction.

Precision: it is a measure of the accuracy of positive predictions made by the model. It is referred as the ratio of true positive (TP) to the total number of data samples, which includes both true positives (TP) and false positives (FP). Precision is mathematically expressed as follows,

$$Pre = \frac{TP}{TP+FP} \quad (23)$$

Where, precision 'Pre' is measured based on the true positive rate (TP), which reflects the number of diabetes diseased samples correctly predicted as diseased, the false positive rate (FP).

Recall:it, also identified as the sensitivity, measures the model's capability to correctly identify positive instances. It is measured as ratio of true positive predictions (TP) to the total number of actual positive instances, including both true positives (TP) and false negatives (FN). Recall is mathematically compute as follows,

$$Rec = \frac{TP}{TP+FN} \quad (24)$$

Where recall 'Rec' called based on the true positive rate (diabetes disease sample correctly classified as diseased) 'TP', FN'indicates a false negative rate.

F1 score: The F1 score represents the harmonic mean of precision and recall, offering a combined metric that balances both performance measures. It is computed using the following formula.

$$F1\ score = 2 * \left[\frac{Pre*Rec}{Pre+Rec}\right] \quad (25)$$

Where, Pre denotes a precision, Rec' indicates a recall .

Specificity: it defines the proportion of actual negative cases that are appropriately identified as negative by the model. Mathematical formula for calculating the specificity is expressed as follows,

$$Specificity = \frac{TN}{TN+FP} \quad (26)$$

Where, the false negative rate (FN), indicating diseased samples incorrectly predicted as non-diseased, and the false positive rate (FP), representing non-diseased samples incorrectly predicted as diseased.

Training time: it is defined as an amount of time consumed during diabetic prediction. The overall training time is calculated as given below.

$$TT = \sum_{i=1}^{m} S_i * Time\ (Res_i) \quad (27)$$

Where, 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 2 comparison of accuracy versus number of data samples

Number of data Samples	Accuracy (%)		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU[2]**
10000	99.1	95.8	97
20000	99.02	95.11	97.11
30000	98.63	95.06	96.85
40000	99.24	95.01	96.78
50000	99.52	94.63	96.32
60000	98.63	94.24	96.75
70000	99.06	94.77	96.52
80000	99.12	94.34	96.87
90000	99.45	94.36	96.36
100000	99.28	94.25	96.42

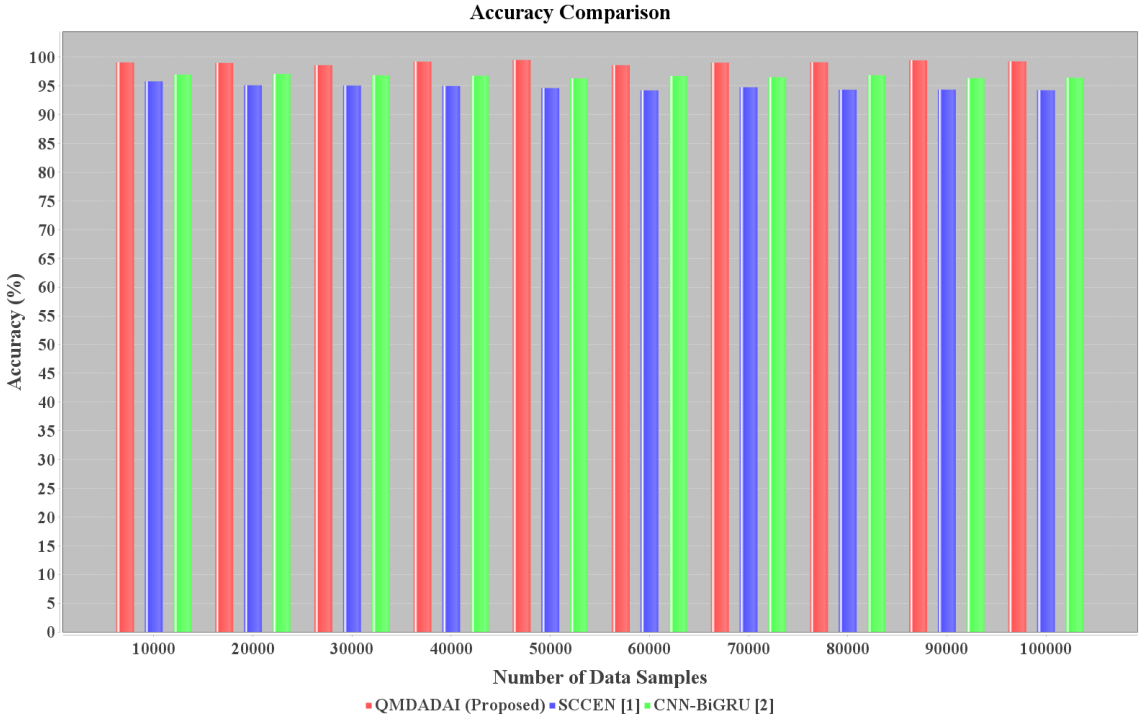


Figure 11 graphical analysis of accuracy

Figure 11 reveals the accuracy of diabetes disease detection versus number of data samples varying from 10000 to 100000, as collected from the kaggle repository dataset. As shown in figure 11, the x-axis indicates the number of sample data, while the vertical axis demonstrates the corresponding accuracy of diabetes disease detection. The graph illustrates that the proposed QMDADAI model outperforms the existing approachesSCCEN [1] and CNN-BiGRU [2]. Let us consider the input data samples of 10000 in the first run. By applying the QMDADAI model, enhanced accuracy of diabetes disease detection is found to be 99.1%. In comparison, the accuracy of the existing methods [1] and [2] are observed to be 95.8% and 97%, respectively.The improved performance is achieved due to the use of asequential rank correlation applied to Discriminative AI model for analyzing both training and testing data in transfer learning model. By utilizing the knowledge obtained from a pre-trained model and adapting it to a new model, the system efficiently handles feature vectors, thereby increasing the accuracy of disease detection. For each technique ten various outcomes are examined and compared. The average of these ten

assessment results illustrate that the accuracy of the QMDADAI model improved considerably by 5% and 2% when compared to existing methods [1] and [2], respectively.

Table 3 comparison of precision versus number of data samples

Number of data Samples	Precision		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
10000	0.993	0.968	0.974
20000	0.991	0.967	0.972
30000	0.986	0.965	0.971
40000	0.99	0.962	0.968
50000	0.987	0.959	0.967
60000	0.985	0.962	0.97
70000	0.984	0.955	0.965
80000	0.983	0.952	0.964
90000	0.985	0.953	0.963
100000	0.988	0.951	0.962

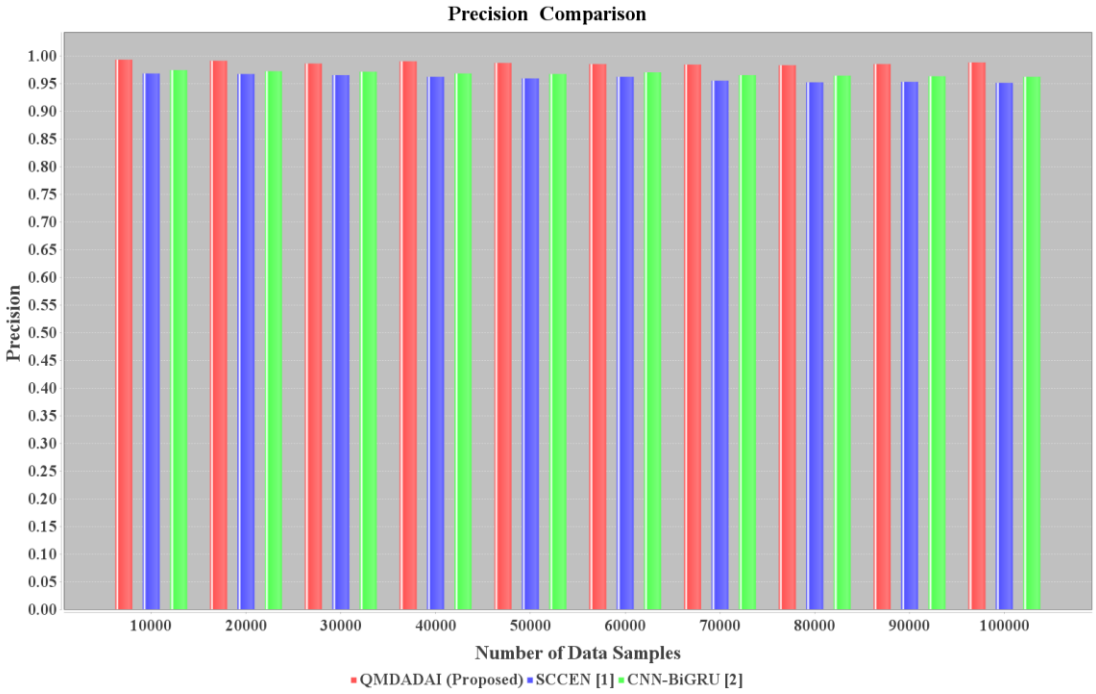


Figure 12 graphical analysis of precision

Figure 12 given above illustrates the graphical outcomes of the precisionin diabetes disease detection using three methods namely QMDADAI model outperforms the existing approachesSCCEN [1] and CNN-BiGRU [2]. In the figure 12, the horizontal axis represents the number of data samples varied from 10000 to 100000, while the vertical axis indicates the precisionoutput. The results display that the precisionperformance of QMDADAI model is improved than that of existing methods [1] and [2]. Considering data samples of 10000 data samples, the QMDADAI model achieved a precision of 0.993. The other two existing methods [1] and [2] observed the precision of 0.968 and 0.974, respectively. . This enhancement is achieved by applying a QMDADAI model to effectively analyze the features in the hidden layer of dense CNN by employing sequential correlation function and provide better disease classification results. Furthermore, the Bats Echolocation optimization is employed to refine the dense CNN layers, aiming to decreasing errors during disease classification. This approach considerably enhances the accuracy by increasing the true positive rate and decreasing false positive in diabetes disease classification, thereby increasing the precision.A comparison of these results reveals that QMDADAI model increases the precision by 3% and 2% when compared to existing methods [1] and [2], respectively.

Table 4 comparison of recall versus number of data samples

Number of data Samples	Recall		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
10000	0.994	0.978	0.987

20000	0.993	0.977	0.986
30000	0.992	0.976	0.985
40000	0.994	0.975	0.984
50000	0.995	0.973	0.983
60000	0.994	0.974	0.982
70000	0.993	0.975	0.984
80000	0.992	0.973	0.983
90000	0.994	0.972	0.985
100000	0.992	0.973	0.984

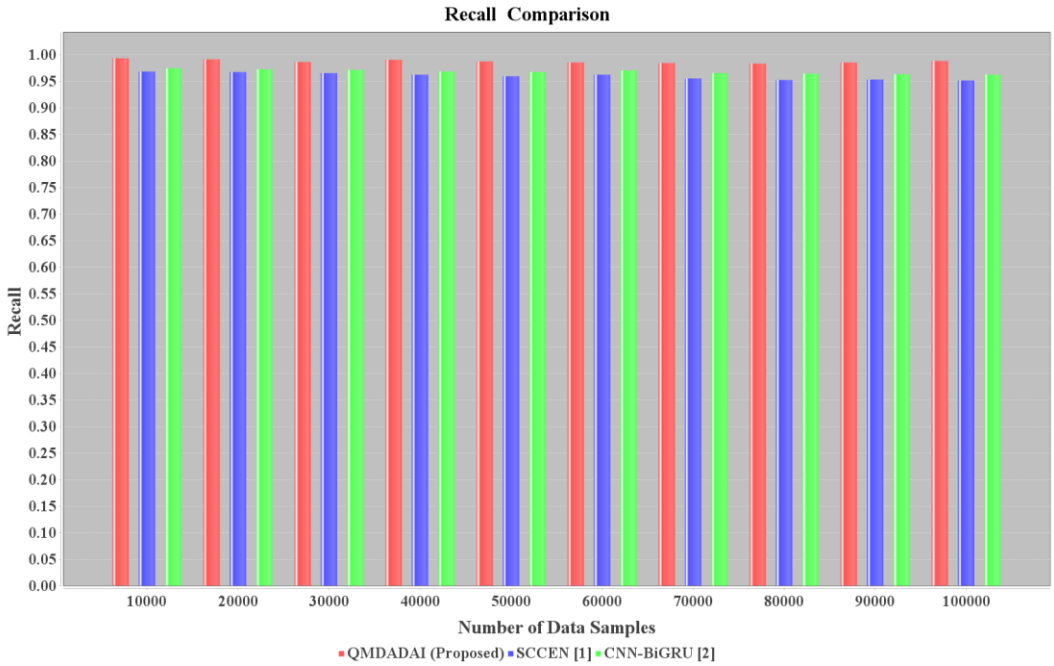


Figure 13 graphical analysis of recall

Figure 13 depicts the performance analysis of recall along with number of data samples ranges from 10000 to 100000 collected from the diabetes disease dataset. The performance of the recall is analyzed by employing three different deep learning methods namely QMDADAI model existing approachesSCCEN [1] and CNN-BiGRU [2]. Among three various methods, the QMDADAI model exhibits a noticeable improvement in recall performance than the existing methods [1] and [2]. For instance, 10000 data samples are considered in first run, the performance analysis of recall using QMDADAI model is found to be 0.994, the recall values of 0.978 and 0.987 are recorded using [1] and [2], respectively. This improvement is achieved by employing robust fine-tuning process in deep transfer learning model. By employing a deep transfer learning approach, the model minimizes the error between predicted and actual classification results through optimal weight detection using Bats Echolocation optimization. This iterative process decreases the false-negative rates and an increase in true positive outcomes for diabetes disease detection. For each deep learning model, various performance results are observed regarding various numbers of input data samples. The overall observed consequences of the QMDADAI model are compared to the existing methods. The comparison results demonstrates that the performance of the recall using QMDADAI model is considerably improved by 2% and 1% compared to [1] and [2], respectively.

Table 5 comparison of F1 score versus number of data samples

Number of data Samples	F1 score		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
10000	0.993	0.972	0.980
20000	0.991	0.971	0.978
30000	0.988	0.970	0.977
40000	0.991	0.968	0.975
50000	0.990	0.965	0.974
60000	0.989	0.967	0.975
70000	0.988	0.964	0.974
80000	0.987	0.962	0.973
90000	0.989	0.962	0.973
100000	0.989	0.961	0.972

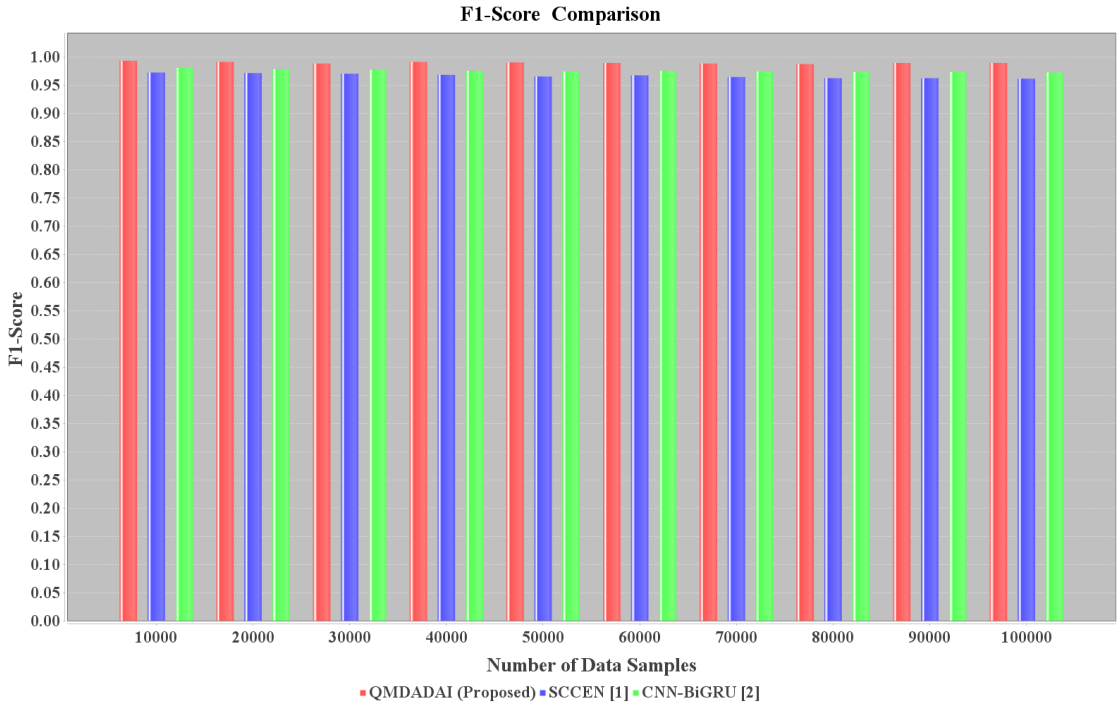


Figure 14 graphical analyses of F1 score

Figure 14 demonstrates the performance analysis of F1-score by varying numbers of data samples ranges from 10000 to 100000 by applying three various methods namely the proposed QMDADAI model existing approachesSCCEN [1] and CNN-BiGRU [2]. The F1-score provides as a harmonic mean of precision in addition to recall, providing a reasonableevaluation of the model efficiency. The enhanced performance of the QMDADAI model is achieved due to the integration of a deep transfer learning policy, which precisely performs the disease classification at the fine tuned output of dense CNN model. Overall, the comparative results emphasize that the QMDADAI model achieves an improvement in F1-score of approximately 2% compared to [1] and 1% compared to [2].

Table 6 comparison of Specificity versus number of data samples

Number of data Samples	Specificity		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
10000	0.977	0.888	0.911
20000	0.974	0.887	0.91
30000	0.973	0.886	0.908
40000	0.972	0.887	0.907
50000	0.97	0.884	0.906
60000	0.971	0.886	0.908
70000	0.973	0.878	0.905
80000	0.974	0.874	0.904
90000	0.972	0.877	0.908
100000	0.974	0.875	0.907

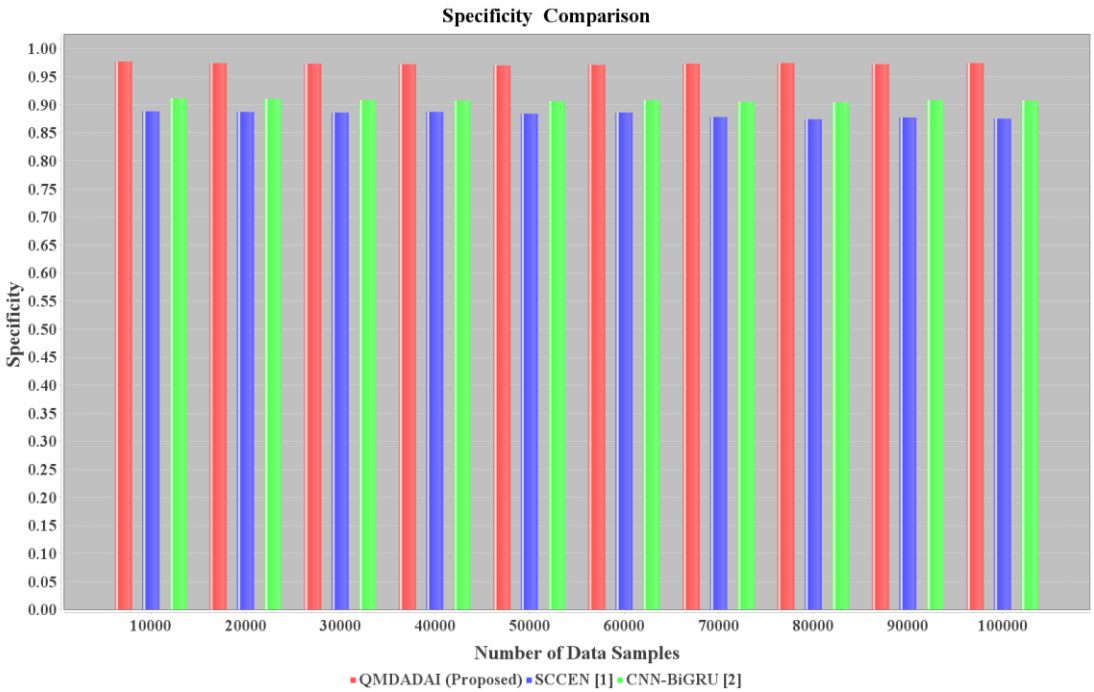


Figure 15 graphical analyses of Specificity

Figure 15 reveal the performance assessment of specificity with respect to the number of data samples, ranging from 10000 to 100000, collected from the diabetes disease dataset. In this graph, horizontal axis indicates the number of data samples, while the vertical axis denotes the corresponding specificity values in diabetes disease classification. The experimental results prove that the proposed QMDADAI model consistently achieved higher specificity than the two existing deep learning models. For example, 10000 data samples are considered in first run to compute the specificity. The QMDADAI model attained a specificity of 0.977, outperforming the existing methods, which observed values of 0.888 and 0.911 respectively. These results are considerably analyzed across different number of data samples to determine the specificity in diabetes disease detection. The comparative analyses illustrates that the QMDADAI model improved specificity by 10% and 7% when compared to [1] and [2].

Table 7 comparison of training time versus number of data samples

Number of data Samples	Training time (sec)		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
10000	250	352	296
20000	274	375	312
30000	296	389	336
40000	310	455	385
50000	332	478	411
60000	359	522	445
70000	412	556	465
80000	445	588	512
90000	475	628	575
100000	511	642	588

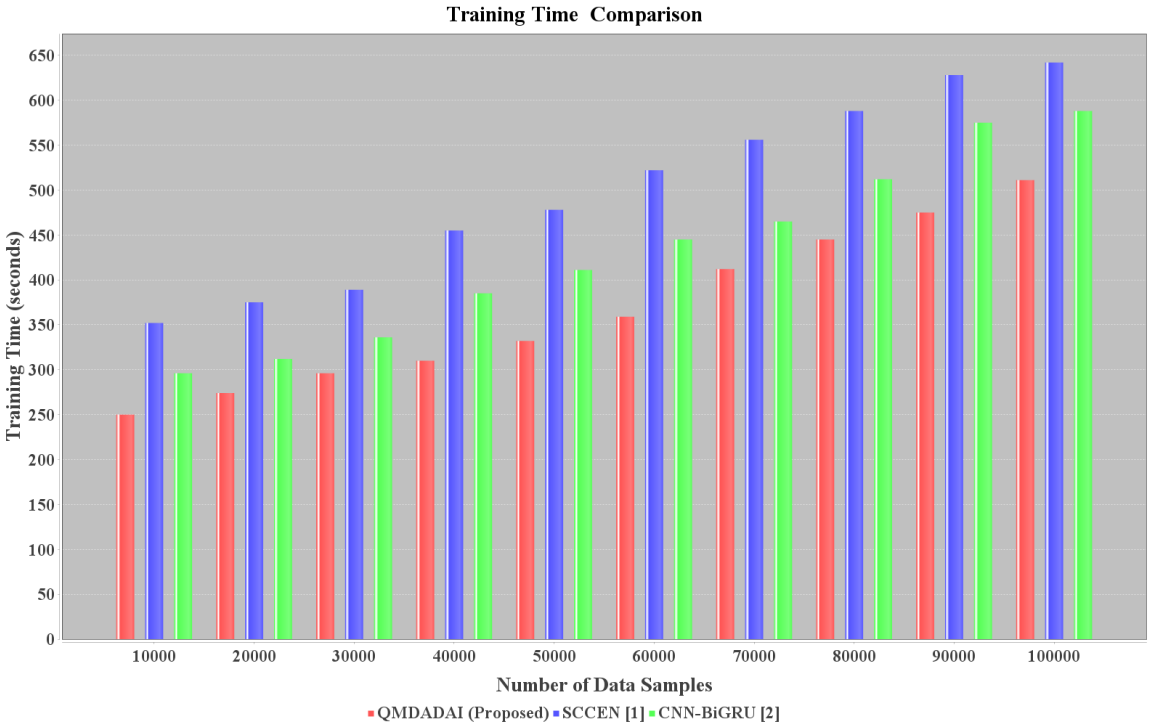


Figure 16 graphical analyses of training time

Figure 16 reveals the overall graphical analysis of training time of diabetes disease classification using three models namely the proposed QMDADAI model and existing approachesSCCEN [1] and CNN-BiGRU [2]. Each model is assessed over ten experimentation runs, using a data of 100000 unique samples. As exposed in the figure 16, training time of all three methods gets increased while enhancing the number of data samples. However, in a specific trial with 10000 samples, the QMDADAI model consumed only 250sec, while [1] and [2] consumed 352sec and 296sec, respectively. This shows that QMDADAI model and achieved a 27% reduction in prediction time compared to [1] and 15% compared to [2]. The effectiveness of the QMDADAI model is achieved owing to its integration of data preprocessing and considerable feature selection method. Specifically, it employs the Mapreduce framework to determine and protect the more significant features while discarding the other features. This proficient reduction in feature space significantlyminimizes the training time of diabetes disease classification.

5.1 Performance analysis of ROC and AUC

The Receiver Operating Characteristic (ROC) curve and the Area under the Curve (AUC) are generally developed to evaluate the performance of disease classification models. The ROC curve represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) across dissimilar threshold values.

Table 8 ROC-AUC results

False positive rate	True positive rate		
	Proposed QMDADAI	**Existing SCCEN [1]**	**Existing CNN-BiGRU [2]**
0	0	0	0
0.1	0.33	0.22	0.26
0.2	0.51	0.33	0.42
0.3	0.67	0.45	0.53
0.4	0.81	0.57	0.67
0.5	0.87	0.66	0.75
0.6	0.92	0.72	0.83
0.7	0.95	0.78	0.88
0.8	0.96	0.85	0.91
0.9	0.97	0.86	0.92
1	0.98	0.93	0.9

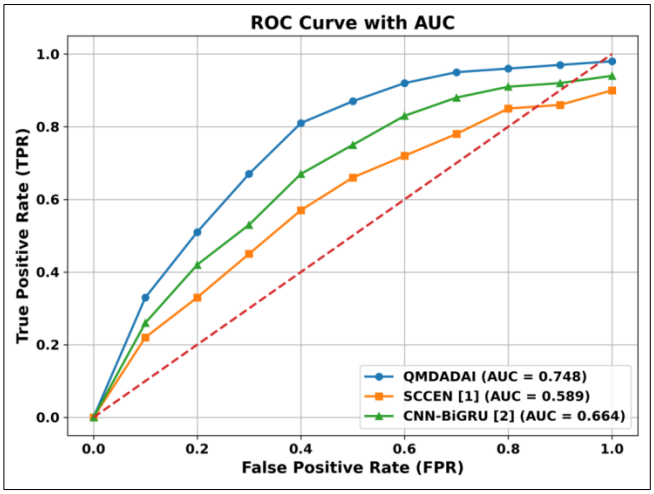


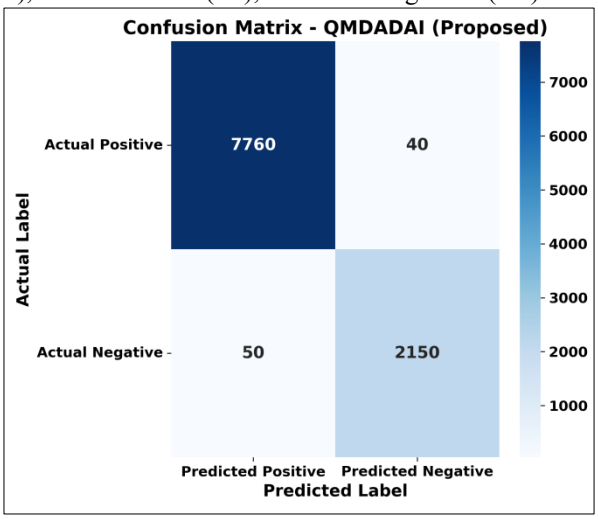
Figure 17 graphical analyses of ROC and AUC

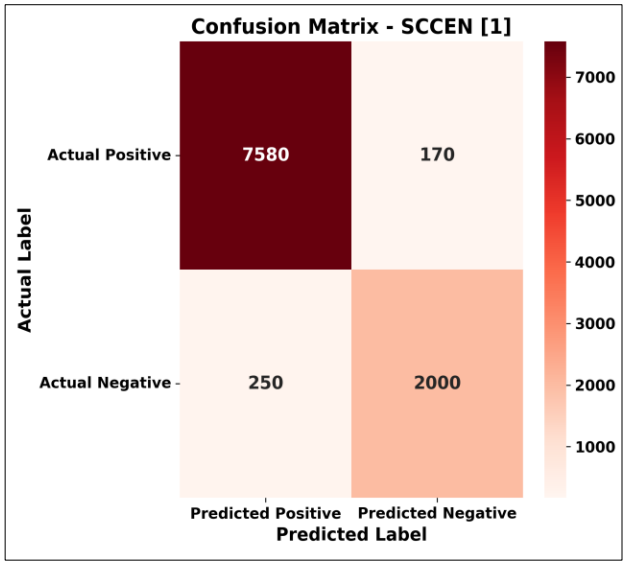
Figure 17 illustrates the ROC–AUC comparison of three classification approaches namely proposed model QMDADAI model and existing approachesSCCEN [1] and CNN-BiGRU [2] evaluated on a dataset of 100000 samples. The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the classification performance of models by plotting the True Positive Rate

(TPR) against the False Positive Rate (FPR) across different threshold values. As observed in the figure, the proposed QMDADAI model consistently achieves an improved TPR for corresponding FPR values when compared to the existing approaches SCCEN [1] and CNN-BiGRU [2] demonstrating its superior performance in diabetes disease classification. The Area under the Curve (AUC) summarizes the overall classification performance into a single value. The AUC ranges from 0 to 1, where a value closer to 1 specifies efficient classification ability, while values below 0.5 reflect poor classification performance. In Figure 17, the diagonal dotted line indicates the random decision boundary, which serves as a threshold point for classification. The obtained AUC scores are 0.748 for the proposed QMDADAI model, 0.589 for [1] and 0.664 for [2]. Although all three deep learning models demonstrate well-organized classification ability, the proposed QMDADAI model achieves the highest AUC, thereby outperforming the existing approaches [1] and [2] in terms of classification accuracy.

5.2 Confusion matrix performance analysis

A confusion matrix is an essential assessment method used to evaluate the performance of diabetes disease classification models, particularly in comparing the proposed QMDADAI model with the SCCEN [1] and CNN-BiGRU [2] using 10,000 data samples. It organizes the diabetes disease classification outcomes into four key categories namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).





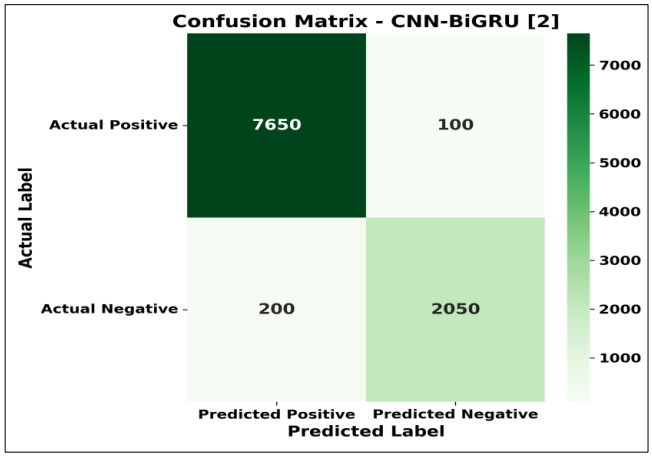


Figure 18 confusion matrix using QMDADAI model and existing SCCEN [1] and CNN-BiGRU [2]

Figure 18 illustrates the confusion matrix of the proposed QMDADAI model based on 10000 data samples. The graphical chart also includes the confusion matrices of the existing approaches, namely SCCEN [1] and CNN-BiGRU [2], enabling a relative performance analysis of their disease classification outcomes. In diabetes disease classification, the confusion matrix provides a comprehensive analysis of model performance by categorizing results into into two classes namely diabetes and no diabetes. It presents the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by each model. The arrangement of these values provides efficient classification outcomes. The QMDADAI model attains a greater number of correctly classified instances (TP and TN) while minimizing incorrect classification (FP and FN), highlighting its improvedresults. Overall, the proposed QMDADAI model exhibits better classification accuracyin comparison with the existing methods.

6. CONCLUSION

Diabetes disease classificationplays a vital role to prevent severe complications. This paper introduces a novel QMDADAI model specifically developed to accurately predicting the diabetes disease. The proposed QMDADAI

model begins with a data preprocessing and more effectiveMapreduce based feature selection aimed to reduce the overall diabetes diseaseprediction time. It also integrates a sequential correlationfunction to analyze efficient samples to enhance diabetes disease classificationaccuracy. To further enhance the model's performance, hyperparameter tuning is employed by Bats Echolocation optimization to decrease the error rates related with diabetes disease classification. Anentireevaluation was conducted using several performance metrics, including accuracy, precision, recall, F1-score, specificity and training timeROC-AUC, confusion matrix across different data samples. The results obtained through quantitative analysis indicate that the QMDADAI model surpasses conventional deep learning approaches by delivering improved accuracy for diabetes disease classification, while also reducing both training time as well as error rates.

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, Springer, Volume 58, 2025, Pages 1-43.https://doi.org/10.1007/s10462-025-11334-3

[2]Tehreem Fatima , Kewen Xia , Wenbiao Yang , Qurat Ul Ain , Poornima Lankani Perera, "Diabetes Prediction Using ADASYN-Based Data Augmentation and CNN-BiGRU Deep Learning Model", Computers, Materials and Continua, Elsevier, Volume 84, Issue 1, 2025, Pages 811-826. https://doi.org/10.32604/cmc.2025.063686

[3] Yu Zhu, Mingxu Zhang, Qinchuan Huang, Xianbo Wu, Li Wan & Ju Huang, "Secretary bird optimization algorithm based on quantum computing and multiple strategies improvement for KELM diabetes classification", Scientific Reports, volume 15, 2025, Pages 1-24.https://doi.org/10.1038/s41598-025-87285-0

[4]Muhammad Sajid, Kaleem Razzaq Malik, Ali Haider Khan, Sajid Iqbal, bdullah A. Alaulamie, Qazi Mudassar Ilyas, "Next-generation diabetes diagnosis and personalized diet-activity management: A hybrid ensemble paradigm", PLoS ONE, Volume 20, Issue 1, Pages 1-22.https://doi.org/10.1371/journal.pone.0307718

[5]Hesham Zaky, Eleni Fthenou, Luma Srour, Thomas Farrell, Mohammed Bashir, Nady El Hajj & Tanvir Alam, "Machine learning based model for the early detection of Gestational Diabetes Mellitus", BMC Medical Informatics and Decision Making, Springer, Volume 25, 2025, Pages 1-15.https://doi.org/10.1186/s12911-025-02947-3

[6]Marwa Matboli, Abdelrahman Khaled, Manar Fouad Ahmed, Manar Yehia Ahmed, Radwa Khaled, Gena M. Elmakromy, Amani Mohamed Abdel Ghani, Marwa M. El-Shafei, Marwa Ramadan M. Abdelhalim & Asmaa Mohamed Abd El Gwad, "Machine learning-based stratification of prediabetes and type 2 diabetes progression", Diabetology & Metabolic Syndrome, Springer, Volume 17, 2025, Pages 1-22.https://doi.org/10.1186/s13098-025-01786-6

[7]Mahmoud Y. Shams, Zahraa Tarek & Ahmed M. Elshewey, "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset", Scientific Reports, volume 15, 2025, Pages 1-22.https://doi.org/10.1038/s41598-024-82420-9

[8]Oyebayo Ridwan Olaniran, Aliu Omotayo Sikiru, JezaAllohibi, Abdulmajeed Atiah Alharbi, and Nada MohammedSaeed Alharbi, "Hybrid Random Feature Selection and Recurrent Neural Network for Diabetes Prediction", Mathematics, Volume 13, Issue 4, 2025, Pages 1-25

[9]Daphne N. Katsarou, Eleni I. Georga, Maria A. Christou, Panagiota A. Christou, Stelios Tigas, Costas Papaloukas& Dimitrios I. Fotiadis, "Optimizing hypoglycaemia prediction in type 1 diabetes with Ensemble Machine Learning modeling", BMC Medical Informatics and Decision Making, Springer, Volume 25, 2025, Pages 1-17.

[10]Jameer Kotwal, Pravin Futane, Gurunath Chavan, Archana Chaudhari, Jithina Jose &Vajid Khan, "Sensor Infused Quantum CNN for Diabetes Disease Prediction and Diet Recommendation", International Journal of Computational Intelligence Systems, Springer, Volume 18, 2025, Pages 1-24

[11]K. Natarajan, Dhanalakshmi Baskaran & Selvakumar Kamalanathan, "An adaptive ensemble feature selection technique for model-agnostic diabetes prediction, Scientific Reports, volume 15, 2025, Pages 1-12.

[12]Agnideep Aich, Md. Monzur Murshed, Sameera Hewage & Amanda Mayeaux, "A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction",Scientific Reports, 2026, Pages 1-26.

[13]Dzira Naufia Jawza, Muhammad Itqan Mazdadi,Andi Farmadi, TriandoHamonangan Saragih, Dwi Kartini and Vugar Abdullayev, "Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature Selection", Journal of electronics electromedical engineering medical informatics, Volume 7, Issue 2, 2025, Pages 295-306.

[14]Sunetra Mukherjee, Debashis De, "XAI-IoT based real-time diabetes prediction using TabTransformer", Medicine in Novel Technology and Devices, Elsevier, Volume 30, March 2026, Pages 1-11.

[15]Radwan Qasrawi, Suliman Thwib, Ghada Issa, Razan Abu Ghoush ,Malak Amro, "Type 2 diabetes risk prediction using glycemic control Metrics: A machine learning approach", Human Nutrition & Metabolism, Elsevier, Volume 42, 2025, Pages 1-9.

[16]Vincent B. Liu, Laura Y. Sue, Oscar Madrid Padilla, Yingnian Wu, "Optimizing blood glucose predictions in type 1 diabetes patients using a stacking ensemble approach", Endocrine and Metabolic Science, Elsevier, Volume 18, June 2025, Pages 1-10.

[17]Siddhant Kumar Jena , Dayal Kumar Behera, Ajay Kumar Jena, Jitendra Kumar Rout, "A Novel Ensemble Machine Learning Framework for Improved Diabetes Prediction and Complication Prevention", Procedia Computer Science, Elsevier, Volume 258 , 2025, Pages 4008-4017.

[18]Sirmayanti, PulungHendro Prastyo, Mahyati, "Enhancing diabetes prediction performance using feature selection based on grey wolf optimizer with autophagy mechanism", Computer Methods and Programs in Biomedicine Update, Elsevier, Volume 8, 2025, Pages 1-16

[19]Pawan Kumar Patidar, Savita Shiwani , Shruti Garg, "Prediction of Diabetes using Machine Learning Classifiers with Polar Bear Optimization", Procedia Computer Science, Elsevier, Volume 258 , 2025, Pages 1338–1347.

[20]Enrico Manzini , Bogdan Vlacho , Josep FranchNadal, Joan Es cudero, Ana Génova, Elisenda Reixach, Erich Andrés , Israel P izarro , Dídac Mauricio, Alexandre Perera-Lluna, "A deep attention-based encoder for the prediction of type 2 diabetes longitudinal outcomes from routinely collected health care data", Expert Systems with Applications, Elsevier, Volume 274, 2025, Pages 1-9.

[21]Rizal, Adelia Fatimahtu Azzahra, Raenault Maxine Salle Bandaso, Jeklin Harefa, Kenny Jingga, "Type 2 Diabetes Prediction Using Recall-Optimized XGBoost", Procedia Computer Science, Elsevier, Volume 269, 2025, Pages 1309-1318.

[22]Niels F. Cleymans, Mark Van De Casteele, Julie Vandewalle, Aster K. Desouter, Frans K. Gorus, Kurt Barbé, "Analyzing Random Forest's Predictive Capability for Type 1 Diabetes Progression", IEEE Open Journal of Instrumentation and Measurement , Volume 4, 2025, Pages 1-11.

[23]V. K. Daliya, T. K. Ramesh, "A Cloud-Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression-

an Azure Machine Learning Perspective", IEEE Access, Volume 13, 2025, Pages 11560 – 11575.

[24]Ugwu Hillary Okwudili, Oparaku Ogbonna Ukachukwu, V. C. Chijindu, Michael Okechukwu Ezea & Buhari Ishaq, "An improved performance model for artificial intelligence-based diabetes prediction", Journal of Electrical Systems and Information Technology, Springer, Volume 12, 2025, Pages 1-30.

[25]Sulaiman Afolabi, Nurudeen Ajadi, Afeez Jimoh, Ibrahim Adenekan, "Predicting diabetes using supervised machine learning algorithms on E-health records", Informatics and Health, Elsevier, Volume 2, Issue 1, March 2025, Pages 9-16.

[26]Yanmin Fan, "Diabetes diagnosis using a hybrid CNN LSTM MLP ensemble", Scientific Reports, volume 15, 2025, Pages 1-15.

[27]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, Volume 14, June 2025, Pages 1-18.

[28]Muhammad Rizwan Khurshid, Sadaf Manzoor, Touseef Sadiq, Lal Hussain, Mohammed Shahbaz Khan Ashit Kumar Dutta, "Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction", PLoS ONE, Volume 20, Issue 1, 2025, Pages 1-29.

[29] Blessing Oluwatobi Olorunfemi, Adewale Opeoluwa Ogunde, Ahmad Almogren, Abidemi Emmanuel Adeniyi, Sunday Adeola Ajagbe, Salil Bharany, Ayman Altameem, Ateeq Ur Rehman, Asif Mehmood & Habib Hamam, "Efficient diagnosis of diabetes mellitus using an improved ensemble method", Scientific Reports, volume 15, 2025, Pages 1-23.

[30]Zahra Rafie, Moslem Sedaghat Talab, Behrooz Ebrahim Zadeh Koor, Ali Garavand, CirruseSalehnasab& Mustafa Ghaderzadeh, "Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes", BMC Public Health, Springer, Volume 25, 2025, Pages 1-17.

