Deep Learning Approach for Type-2 Diabetes Prediction from Clinical Data

Ph.D. Synopsis Submitted to

Gujarat Technological University for

The Degree of

Doctor of Philosophy In

Computer/IT Engineering

By

HITESH B PATEL

Enrollment No: 199999913574

Supervisor:

Dr. Keyur N Brahmbhatt, Professor & Head (I.T. Dept.)

BVM Engineering College, VV Nagar, Anand, Gujarat.

INDEX

1. Title of the Thesis and Abstract

1.1 Title of the Thesis

Deep Learning Approach for Type-2 Diabetes Prediction from Clinical Data

1.2 Abstract

About half of the world's population suffers from diabetes mellitus. In most cases, type 2 diabetes is the most widely perceived typecast. It is defined by hyperglycemia, or unusually high blood glucose levels, and is frequently accompanied by a variety of consequences [1]–[5]. However, many people use the same glucose-lowering medicine. Current type 2 diabetes recommendations advice selecting glucose-lowering treatment options based on clinical factors which are compatible with the primary objective of precision medicine: personalizing medical therapy to an individual. Type 2 diabetes is affected by both genetic and non-genetic factors, including hyperglycemia, insulin resistance, and decreased insulin production.

The enormous development in technology makes diagnosing easier in the medical field under various existing approaches. One significant drawback pertained to the challenge of reducing disease treatment costs effectively. The research revealed that certain network features, which play a crucial role in decision-making processes, were characterized by low accuracy values. This finding implies that although these features are integral to the diagnostic process, their predictive power was not sufficiently reliable. As an outcome, the performance accuracy of the overall prognosis prototype was compromised, leading to potential inaccuracies in identifying and classifying diabetic cases at an early stage. These limitations underscore the importance of addressing the accuracy of critical network features to enhance the efficacy of diabetic prognosis models and ultimately optimize disease management strategies. A new deep-learning technique for diabetes detection is proposed in this research that resolves the challenges in the existing.

The proposed Algorithm-1 for Type-2 diabetes prediction and classification is developed using the publicly available PIMA Diabetes Dataset. The framework incorporates several preprocessing steps, including missing value handling, outlier detection, and data standardization. To address class imbalance, a class-wise data augmentation strategy is employed using an Adversarial Variational Auto-Encoder (AVAE), followed by classification using a Random Forest model. The performance of the proposed approach is experimentally evaluated and compared with multiple machine learning techniques, namely Logistic Regression, Decision Tree, AdaBoost, Support Vector Classifier (SVC), Random Forest, Gradient Boosting, and K-Nearest Neighbors, using cross-validation based on mean accuracy. The results indicate that the Random Forest classifier combined with class-wise data augmentation achieves a prediction accuracy of 93.09% for Type-2 diabetes detection.

The approach for proposed algorithm-2 combines Adversarial Variational Auto-Encoder (AVAE) for data augmentation with a Morlet Wavelet-assisted Deep Learning Network featuring Fast Fourier Transform (FFT) Overlap and Add Convolution (MW-FFT-OAconv) to enhance classification accuracy for PIMA Diabetes Dataset. A new optimization technique, referred to as the Weighted Mean of Vectors (WMOV), is proposed to

estimate the weight parameters of the MW-FFT-OAconv network. The effectiveness of the proposed method is validated through experimental evaluation using several statistical performance metrics, including accuracy score, F1-measure, Precision; True negative and true positive rate (TNR & TPR), Error Percentage of Classifier (CEP), and Matthews correlation coefficient (MCC). With respect to machine learning prediction models (naive Bayes, random forest, and decision tree), the proposed technique achieves an accuracy level of 98.43% in predicting type 2 diabetes.

2 Brief description of the state of the art of the Research Topic

Type 2 diabetes mellitus (T2DM) is among the most prevalent metabolic disorders worldwide. It develops when the pancreas does not secrete sufficient insulin or when the body becomes ineffective in utilizing the insulin produced. Proper regulation of insulin secretion and action is essential to satisfy metabolic requirements. Owing to its long-term nature and widespread occurrence, Type II diabetes represents a major global public health concern. The rising global prevalence, projected to affect 592 million people by 2035, underscores the urgency of effective management and prevention strategies [6]. Effective early prediction and intervention are critical in mitigating the disease's progression and associated complications. Advances in machine learning have been pivotal in predicting diabetes onset and complications, enabling tailored preventive measures and better patient outcomes [7], [8].

Reza MS et al.[9], explores the use of Support Vector Machines (SVM) for Type II diabetes prediction, introducing a novel integrated kernel framework that combines RBF City Block kernels and Radial Basis Function (RBF). Leveraging PIMA dataset, the proposed model addresses data challenges such as class imbalance, missing values, and outliers. The results demonstrate an impressive accuracy score of 85.5%, showcasing the model's better performance than existing methods. These findings emphasize the potential of the proposed approach in aiding clinical decision-making for early diabetes detection.

Kibria HB et al.[10], introduces an ensemble approach for diabetes prediction using soft voting based classifier, integrating six algorithms of machine learning, including XGBoost, Random Forest, along with explainable AI tools like SHAP. Employing the PIMA Indian Diabetes dataset, the model applies advanced preprocessing techniques such as SMOTETomek for class balancing and median imputation for missing values. A 90% balanced accuracy with 89% F1 score, the presented model not only outperforms existing methods but also enhances interpretability for clinical application, aiding in early-stage diabetes detection.

Sampath P et al.[11], presents a robust framework for diabetes prediction by integrating preprocessing techniques like SMOTE for class balancing, outlier rejection, and feature selection with ensemble-based machine learning models. Using the Diabetes dataset of Pima Indian, the study applies an ensemble of AdaBoost and XGBoost, optimized via grid search. The results demonstrate exceptional performance, achieving 90.4% accuracy and 0.968 AUC, surpassing current futuristic methods. These research discoveries emphasize the model's potential in enhancing diabetes diagnosis and healthcare outcomes.

Albadri RF et al.[12], introduces a novel hybrid machine learning framework that combines Support Vector

Machine (SVM), Decision Tree (DT), and Random Forest (RF) classifiers to improve prediction accuracy. Using the Pima Indian Diabetes dataset and key clinical attributes such as glucose level, BMI, age, and insulin, the proposed model demonstrates strong predictive capability. Rigorous evaluation using the holdout method and k-fold cross-validation results in accuracies of 88.5% and 90.1%, respectively. These outcomes indicate the robustness of the approach and its potential applicability in enhancing clinical decision-making for diabetes risk prediction.

García-Ordás MT et al.[13], proposed a deep learning–based pipeline and was introduced for predicting diabetes, incorporating multiple stages to enhance model performance. The framework applies data augmentation through a Variational Autoencoder (VAE), feature enhancement using a Sparse Autoencoder (SAE), and classification via a Convolutional Neural Network (CNN). The approach was evaluated on the Pima Indians Diabetes Database, which includes patient attributes such as number of pregnancies, glucose and insulin levels, blood pressure, and age. Experimental results show that training the CNN jointly with the SAE on a well-balanced dataset achieves an accuracy of 92.31%.

Shao et al. [14], proposed a modified stacked autoencoder (MSAE) incorporating an adaptive Morlet wavelet was introduced for the automatic diagnosis of different fault types and severity levels in rotating machinery. In this approach, the Morlet wavelet activation function is employed to build the MSAE, enabling an effective nonlinear mapping between raw non-stationary vibration signals and corresponding fault conditions. To further enhance performance, a nonnegative constraint is integrated into the cost function, improving both sparsity and reconstruction accuracy. The proposed technique is validated using raw vibration data obtained from a sun gear unit and a roller bearing unit. Experimental results demonstrate that the method outperforms several existing state-of-the-art fault diagnosis approaches.

Chitsaz et al. [15], proposed a new approach for processing convolutional neural networks in the FFT domain based on an input-splitting strategy. The method addresses the limitations associated with computing FFTs using small kernels, which commonly arise in CNN applications. By employing input splitting as an effective solution, the need for redundancy techniques such as overlap-and-add is reduced, leading to improved computational efficiency.

The study of Ahmadianfar et al.[16], introduced a detailed study on a novel optimization algorithm called the Weighted Mean of Vectors (INFO), focusing on its underlying concepts and analytical framework for solving diverse optimization problems. INFO is an enhanced version of the traditional weighted mean approach, where the weighted averaging concept is utilized to construct a robust optimization mechanism. The algorithm updates candidate solution vectors through three main components: a position update strategy, vector combination, and a local search process. Findings reported in the literature show that INFO achieves superior performance compared to both conventional and advanced optimization techniques, particularly in balancing exploration and exploitation. For engineering optimization tasks, INFO demonstrates strong convergence capability, achieving solutions within 0.99% of the global optimum.

Kumar P et al.[17], proposed a Deep Neural Network (DNN)–based classifier introduced for diabetes

prediction using an unsupervised learning framework applied to the Pima Indian Diabetes (PID) dataset. Feature selection was carried out using a feature importance mechanism that integrates Extra Trees and Random Forest algorithms. The PID dataset was sourced from the UCI Machine Learning Repository, and multiple train–test split configurations were explored during experimentation. Model performance was assessed using standard evaluation metrics, including accuracy, specificity, sensitivity, recall, and precision. The proposed approach achieved a classification accuracy of 98.16% with a random train–test split, demonstrating superior performance compared to several existing state-of-the-art techniques.

Using a Machine learning (ML) technique, Chang et al. [18], developed an electronic diagnostic framework and was designed to support the detection of diabetes mellitus, specifically type 2 diabetes, within an Internet of Medical Things (IoMT) environment. One of the major challenges in healthcare-oriented machine learning systems is their limited transparency in explaining decision-making processes, which often reduces user trust and acceptance. To address this issue, three interpretable supervised learning algorithms—Naive Bayes, Random Forest, and J48 decision tree—were developed and evaluated using the Pima Indians Diabetes dataset. In addition to performance evaluation, the decision-making behavior of each model was analyzed to improve interpretability. Experimental results indicate that the Naive Bayes classifier is well suited for binary classification tasks with effective feature selection, while the Random Forest model achieves better performance when handling a larger number of features.

Omisore et al.[19] , introduced a new effective learning model to detect diabetes mellitus with personalized management. A multimodal adaptive neuro-fuzzy inference model is used to identify diabetes as part of the integrated system, and a knowledge-based dietary recommendation model is utilized to treat diabetes based on an individualized diet. The Pima Indians diabetes dataset was used for training and validation, and the publicly accessible Schorling diabetes dataset was used for revalidation. A dataset around 14 samples of female patients from the Obafemi Awolowo University Hospital Complex in Ile-Ife, Nigeria, was used to apply the model retrospectively. Furthermore, the recommendation algorithm used the results of users' eating formulas with their diagnoses to forecast users' daily food intake and generate weekly customized meal plans based on a professionally created template.

3 Definition of the problem

Diabetes prediction is a complex problem since the class distributions across the features are not linearly separable. In the last few decades, several techniques have been introduced to predict the diabetes. The prediction task can be carried out using various techniques, such as AdaBoost, Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), logistic regression, and Convolutional Neural Networks (CNN). The sensitivity of AdaBoost classifier to noisy input is high as faults in one stump impact other faults which can extent within the forest of stumps. The computational complexity of CNN is high as compared to logistic regression; however, it is suitable only for linearly separable input.

Deep learning models, particularly CNNs, often achieve the highest levels of prediction accuracy. Moreover, convolutional layers in CNNs typically require large volumes of input data, whereas datasets for Type 2

diabetes are often limited in size. Therefore, increasing the data dimensionality becomes essential to enable more effective training of CNN models. Additionally, enhancing deep feature representations through feature augmentation can further improve the classification accuracy of the network. However, no report on successfully improving the classification accuracy of Type 2 Diabetes with deep features augmentation has been published. Furthermore, if the CNN used general activation functions then it will not establish exact mapping between a nonlinear input data and several output patterns. The computation complexity of CNN is mainly depends on convolution layer. As a result of this, the execution time of the Type 2 Diabetes prediction model will be increased. These points motivate us to propose a new Type 2 Diabetes prediction model based on deep feature augmentation and simplified CNN classifier.

4 Objective & Scope of the Research

4.1 Objective

The primary aim of this study is to design a deep learning–based framework for the accurate diagnosis of diabetes mellitus. The key contributions of this work are summarized as follows:

- To perform data augmentation both in samples and features by combining the benefits of Variational Auto-Encoder for data generation and capability of the adversarial learning for the latent representation. This augmentation will be used for retaining the valuable information of original data.
- To improve the mapping ability of the classifier between the augmented features and different Diabetes states by replacing the activation function of the hidden layer with Morlet wavelet function.
- To minimize the computational cost and training time of the deep learning classifier by introducing a novel FFT-based convolution approach.
- To improve the accuracy of deep neural network by selecting optimal hyper parameters using metaheuristic approach.

4.2 Scope

This research focuses on the following major aspects:

1. To investigate and identify an effective data augmentation strategy for generating synthetic samples.
2. To study and choose suitable hidden-layer activation function that provides optimal feature mapping between augmented data and various diabetes conditions.
3. To explore and adopt efficient techniques for reducing the computational cost and execution time of the deep learning classifier.
4. To examine and select an optimal metaheuristic optimization method for enhancing the accuracy of the deep neural network through optimal hyperparameter tuning.
5. Develop a Type 2 Diabetes Prediction model with Advanced Data Augmentation technique using AVAE with Random Forest Classifier using PIMA Diabetes Dataset and Diabetes Prediction Dataset.

6. Implement the model for Type 2 Diabetes Prediction based on feature augmentation using AVAE and Morlet Wavelet assisted Deep Neural Network with FFT Overlap and Add Convolution and Weighted Mean of Vectors for optimization (MW-FFT-OAconv) using PIMA Diabetes Dataset.

By achieving these objectives within the specified scope, the research work aims to develop a new Type 2 Diabetes prediction model based on deep feature augmentation and simplified classifier.

5 Original contributions by the Thesis

The contributions of a thesis aimed at improving the performance of a Type 2 diabetes prediction model through data augmentation can be significant in multiple aspects:

1. In this thesis different data augmentation techniques are evaluated to generate synthetic data.
2. A novel data augmentation framework combining AVAE and SMOTE with a Random Forest classifier is proposed for Type 2 diabetes prediction, leading to improved performance on the PIMA Diabetes Dataset and an additional Diabetes Prediction dataset. The effectiveness of the approach is validated using statistical metrics such as Accuracy, Precision, Classifier Error Rate, F1-Score, and Matthews Correlation Coefficient (MCC).
3. Integrated AVAE with Morlet Wavelet assisted Deep Neural Network with FFT Overlap and Add Convolution and Weighted Mean of Vectors for optimization to further improve performance of proposed Type 2 Diabetes Prediction model.
4. The proposed integrated framework is tested on the PIMA Diabetes Dataset and its performance is systematically compared with conventional machine learning models as well as deep neural network–based approaches.

6 Methodologies of the Research / Results - Comparison

For Type 2 diabetes prediction, a benchmark dataset—the Pima Indians Diabetes Dataset—is utilized, which is publicly accessible through Kaggle at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. The PIMA dataset represents a specific demographic—Native American women—which limits its applicability to broader populations. The dataset's small size also poses challenges for training deep learning models.

The Type 2 diabetes prediction in this study is based on the widely used Pima Indians Diabetes Dataset, which contains 768 instances described by eight independent attributes: pregnancies, skin thickness, blood pressure, glucose level, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable represents two classes: non-diabetic (0) and diabetic (1). Among the total samples, 500 records belong to the non-diabetic class, while 268 correspond to the diabetic class, indicating class imbalance.

Several features in the dataset—such as glucose level, insulin level, blood pressure, skin thickness and BMI—contain zero values that are physiologically implausible. To address this issue, these zero entries were replaced with the mean values of their respective attributes.

Outlier handling was performed using the Interquartile Range (IQR) method, which is a robust statistical technique for identifying extreme values outside the normal data distribution. Outliers were examined in features including skin thickness, pregnancies, glucose and insulin levels, blood pressure, BMI, age, and diabetes pedigree function. For the purpose of data augmentation, detected outliers were removed to improve data quality and model reliability.

6.1 Data Pre-Processing

6.1.1 Replacing/Filling incorrect values

Filling in the missing, incorrect or null values comes after removing the outliers, which can cause any classifier to forecast incorrectly. Equation (1) indicates that, in the proposed framework, missing or invalid attribute values are handled by imputing them with the corresponding mean values instead of discarding the data samples. Mean imputation was chosen to handle missing values due to its computational simplicity and effectiveness in datasets with normally distributed features. Alternatives such as k-NN imputation were considered but it is computationally expensive for the given dataset as size of the dataset is limited.

$$I(x) = \begin{cases} \text{mean}(x), & \text{if } x = 0/\text{null} \\ x, & \text{otherwise} \end{cases} \quad (1)$$

6.1.2 Outlier Rejection

Classifiers are highly sensitive to the range and distribution of feature values; therefore, anomalous data must either be removed or appropriately treated to preserve the integrity of the dataset. As expressed in Equation (2), this study formulates a mathematical criterion for identifying and eliminating outliers from the data distribution.

$$L(x) = \begin{cases} x, & Q1 - 1.5 \times IQR \leq X \leq Q3 + 1.5 \times IQR \\ \text{reject}, & \text{otherwise} \end{cases} \quad (2)$$

It lies in an n-dimensional space, where the feature vector is represented as $x \in J^n$. Here, Q1, Q3 and IQR denote the first quartile, third quartile, and interquartile range, respectively, with, Q1, Q3, IQR $\in J^n$.

For outlier detection, the Interquartile Range (IQR) method was selected for its robustness to non-normal distributions. While techniques like Z-score standardization and Hampel filtering could also be effective, IQR aligns well with the dataset's characteristics. Future work could compare these methods experimentally to optimize preprocessing further.

6.1.3 Standardization

Data scaling is used to normalize the range of data belongings. Standardization transforms variables into a common proportion. The process of standardization allows us to distinguish scores between dissimilar types of variables that may have contrasting original units or scales.

The formula for standardization of a data point value x for a given dataset with mean (μ) and standard deviation (σ) is:

$$z = \frac{x-\sigma}{\mu} \quad (3)$$

Where z is the standardize value.

6.2 Data Augmentation using AVAE

Adversarial Variational Auto Encoder combines the benefits of both principles of Variational Auto Encoders and Generative Adversarial Networks. For compact representation of data values VAEs work fine and for generation of new samples GANs are more proficient. AVAE integrates these capabilities, allowing it to generate new data samples that are similar to the training dataset samples.

The AVAE architecture consists of a three-layer encoder and decoder with 64, 128, and 256 neurons, respectively, using ReLU activation. The adversarial network comprises two dense layers with sigmoid activation, optimizing a combined loss function of reconstruction loss and adversarial loss. The model was trained using the Adam optimizer with a learning rate of 0.001. To mitigate overfitting, an early stopping strategy was applied with a patience of 10 epochs. The suggested model's AVAE for data augmentation is shown in figure 1.

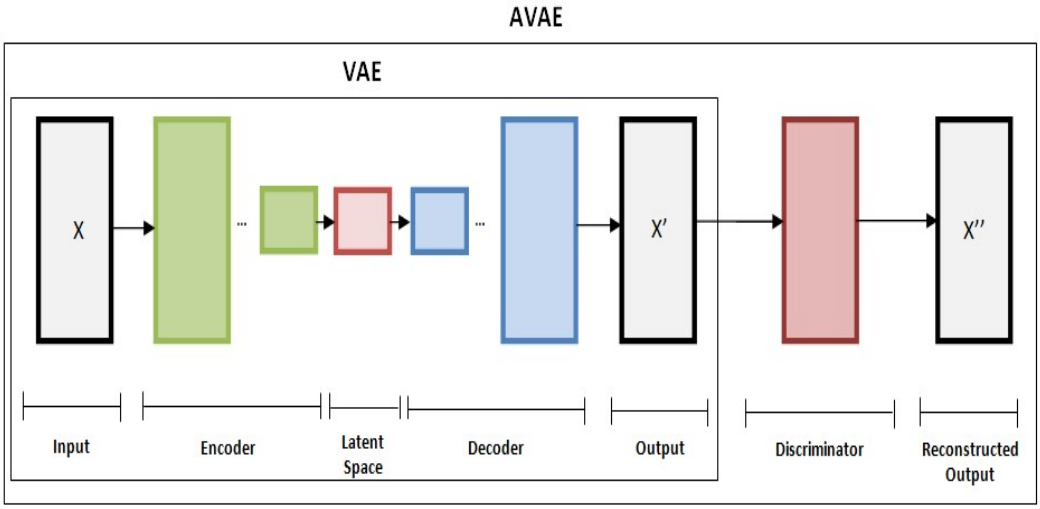

Figure-1 Adversarial Variational Auto Encoder (AVAE) Architecture

6.3 Proposed Methodology using Random Forest Classifier

After, data pre-processing, an Adversarial Variational Auto-encoder is used to expand the amount of data for diabetic and non diabetic class separately because after augmentation we can easily set target features according to which class they are belonging to. Finally, Random Forest Classifiers is used to classify normal patients and type 2 diabetic patients. Figure 2 displays the architectural representation of proposed methodology.

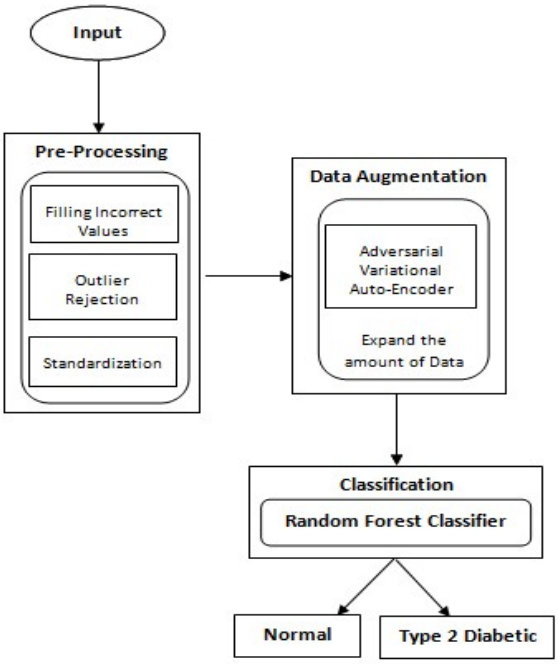

Figure-2 Flow diagram of data augmentation based type 2 diabetes prediction

6.3.1 Random Forest Classifier

A Random Forest Classifier is an ensemble learning algorithm that builds multiple decision trees during training and combines their predictions to produce a final classification result, typically through majority voting. Each tree is trained on a random subset of the data and a random subset of features, which helps reduce overfitting and improves generalization.

Random Forests are robust to noise, and can handle high-dimensional data. Additionally, they provide measures of feature importance, making them useful for both prediction and interpretability in classification tasks. This classifier handles non-linear relationships effectively and helps to reduces overfitting compared to a single decision tree and also it can estimate feature importance as well.

6.3.2 Results & Discussions

Unlike traditional techniques like VAE, SMOTE and GAN used in literature, our approach introduced Adversarial Variational Auto Encoder (AVAE) for data augmentation. This method not only balances the dataset but also generates synthetic samples with greater diversity and realism, improving model generalizability. Furthermore, this study highlights the distinctive synergy between AVAE-based data augmentation and the Random Forest classifier, exploiting Random Forest's robustness to noise and its effectiveness in capturing complex feature interactions. The integration of adversarial approaches offers additional resilience to model overfitting, addressing limitations of previous models that often lacked this feature.

Following is the comparison of accuracy score achieved by our proposed methodology with relevant literature. All studies which uses PIMA diabetes dataset in terms of accuracy.

Research Work	Adopted Methods	Accuracy Score
Reza MS et al. [9]	SMOTE, SVM with Proposed Integrated Kernel	85.50 %
Kibria HB et al. [10]	SMOTETomek, XGB + RF	90.00 %
Sampath P et al. [11]	SMOTE, AdaBoost + XGBoost	90.40 %
Albadri RF et at. [12]	Hybrid ML Algorithm	88.50 %
García-Ordás MT et al. [13]	VAE with CNN	92.31 %
Al-Dabbasa L et al. [20]	SVMSMOTE , XGBoost	91.00 %
Ganie SM et al. [21]	Gradient boosting	92.85 %
Ahamed BS et al. [22]	LightGBM	92.50 %
Moreno-Barea FJ et al. [23]	VAE, GAN, Logistic Regression (with 50% Data Augmentation)	77.95 %
Sabitha E et al. [24]	SMOTE, RF	81.25 %
This Study	AVAE, Random Forest	93.08 %

Table 1: Comparison of proposed work with various research work

Our study outperforms existing research by achieving the highest accuracy of 93.08%. The innovative use of Adversarial Variational Auto Encoder (AVAE) for data augmentation introduces realistic and diverse synthetic samples than traditional data augmentation methods. Combining AVAE with Random Forest enhances predictive robustness and interpretability, providing insights into key risk factors. This novel hybrid strategy enhances predictive accuracy while also improving generalizability and clinical applicability, thereby establishing a new benchmark for diabetes prediction research.

Figure 3 illustrates the performance of the proposed approach in terms of Accuracy Score, F-Measure, Precision, and Recall. When compared with Logistic Regression, Decision Tree, and Naive Bayes classifiers, the proposed method achieves superior results, attaining the highest accuracy of 93.09%.

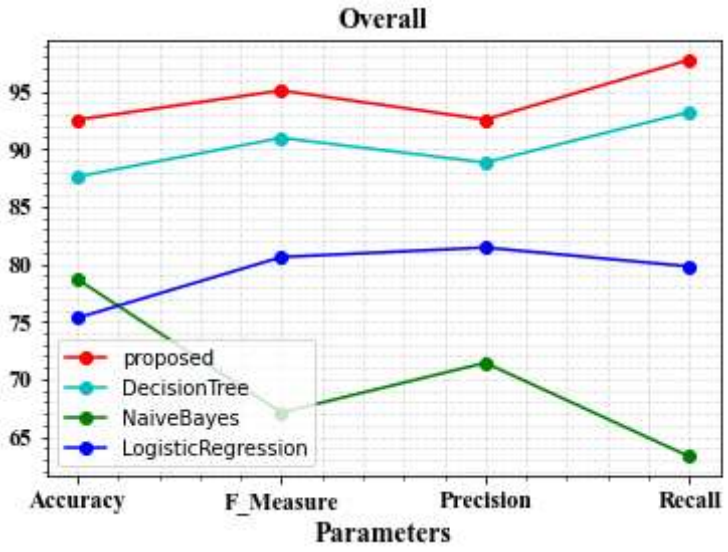

Figure-3 Comparison Graph of Accuracy, F-Measure, Precision, Recall

Underperformance of Logistic Regression and Decision Trees struggled because of their limited ability to model non-linear relationships. Techniques like SMOTE generated synthetic samples but lacked diversity, reducing the effectiveness of classifiers like Random Forest. In contrast, AVAE produced realistic and diverse synthetic data, improving feature representation and overall model performance.

This study introduces a novel diabetes prediction framework combining Adversarial Variational Auto Encoder (AVAE) for data augmentation and Random Forest for classification. The study's quantitative results highlight AVAE's transformative potential in healthcare analytics. Strengths include superior accuracy, enhanced data diversity, and interpretability of feature importance.

6.4 Proposed Methodology based on Customized Classifier (MW-FFT-OAconv)

The pre-processing stage includes outlier rejection, filling in missing values and normalization. Initially, the input data that are deviated from other observations are rejected by formulating a mathematical equation based on different quartile ranges of the attributes.

Subsequently, missing or null values in the dataset are imputed using the mean values of the corresponding attributes. The data is then preprocessed using a normalization technique to scale each feature to a comparable range within the set of real numbers. To increase the amount of data and features in the less representative classes, an AVAE is developed. In this approach, more meaningful information is extracted from raw data by learning the latent distribution of the original training samples.

The flow diagram of type 2 diabetes based on feature augmentation and Morlet wavelet-assisted deep learning network with FFT overlap and adds convolution is shown in Figure 4.

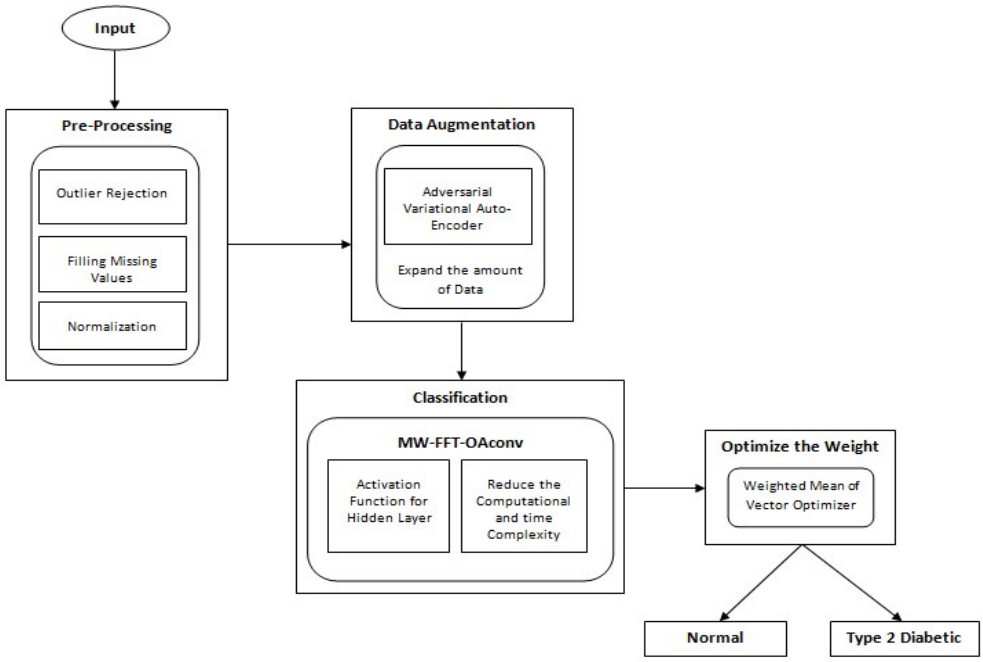

Figure-4 Flow diagram of type 2 diabetes based on data augmentation and Morlet Wavelet assisted deep learning network with FFT overlap and add convolution.

The proposed MW-FFT-OAconv classifier is used to identify type 2 diabetes. The Morlet wavelet activation function is employed to establish an accurate nonlinear mapping between the augmented feature representations and the different stages of diabetes for the proposed MW-FFT-OAconv classifier. In addition, to reduce the computational and time complexity of training and interface, FFT overlap and add convolution will be replaced by deep learning algorithm convolution layers. Finally, the MW-FFT-OAconv architecture parameters (weight) will be employed to generate a novel weighted mean vector optimizer.

6.4.1 Morlet wavelet

It is a new activation function for neural networks that is based on wavelet analysis. It effectively exploits time–frequency localization through the wavelet transform framework.

By using Morlet wavelets activation function in the hidden layer of the basic autoencoder, it is possible to build an accurate nonlinear mapping between the augmented features and different diabetes states because the characteristics are more comparable to one another.

The Morlet wavelet is transmitted as:

$$\varphi(\mathrm{h}) = \frac{1}{\sqrt{f_b}} \cos(2\pi f_c h)\exp(-h/f_b) \qquad (4)$$

Here, the performance of the Morlet wavelet function is strongly influenced by the selection of the parameters fb and fc. Also, fb and fc denotes the central frequency and bandwidth, respectively.

6.4.2 FFT overlap and add convolution

FFT Overlap–Add convolution computes linear convolution efficiently by processing the input signal in blocks. Initially, the signal is segmented into uniform blocks. To prevent circular convolution effects, each segment is extended with zeros so that its length is no less than the combined size of the block and the impulse response minus one. The FFT of the zero-padded block is then computed and multiplied with the FFT of the impulse response in the frequency domain. The transformed result is then brought back into the time domain using an inverse FFT, yielding an intermediate output segment. Because each block convolution produces overlapping samples, the overlapping portions of successive output blocks are added together to form the final output signal.

The key points are that Overlap–Add produces the same result as linear convolution, it significantly reduces computational complexity from O(N2) to O(NlogN) for long signals, it is well suited for large or streaming data, and it relies on FFT, IFFT, zero-padding, and overlap-and-add operations to eliminate circular convolution effects.

6.4.3 Weighted Mean of Vectors Optimization

The architecture parameters (e.g., weight) of MW-FFT-OAconv are obtained using a new weighted mean of the vector optimizer. The vector optimization algorithm's weighted mean is used here to increase classification accuracy by optimizing the weight parameter. This study introduces the optimization algorithm designed by Ahmadianfar et al. [16].

The Weighted Mean of Vectors Optimizer is a population-driven metaheuristic method that refines candidate solutions by guiding them toward a centrally weighted position derived from higher-quality solutions within the search space. Instead of relying on gradients, Weighted Mean of Vectors Optimizer evaluates multiple solution vectors, assigns higher weights to better-performing vectors (based on fitness), and computes a weighted average that represents a promising search direction.

Weighted Mean of Vectors Optimizer is simple, flexible, and suitable for nonlinear, non-convex, and complex optimization problems where derivative information is unavailable.

6.4.4 Results & Discussions

The experiment used the freely available PIMA dataset. The proposed framework is developed using Python, and its performance is experimentally assessed and benchmarked against existing machine learning prediction models using statistical evaluation metrics such as CEP, F1-score, and MCC.

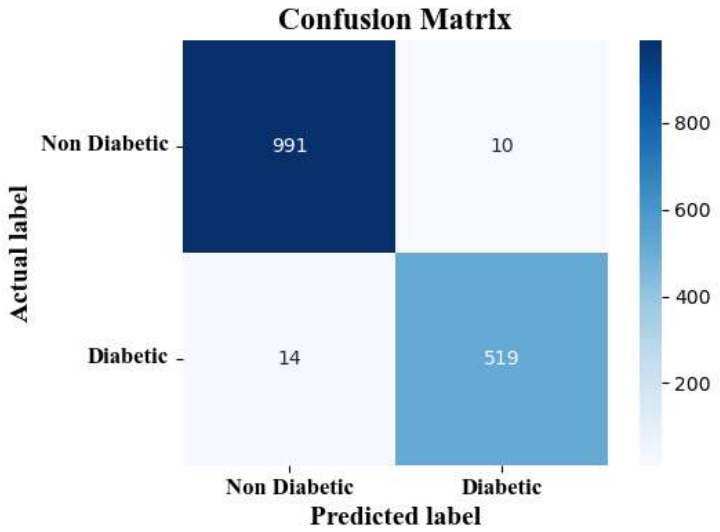

Figure-5 Confusion Matrix

Figure 5 presents the confusion matrix illustrating the correspondence between the predicted outcomes and the actual class labels for the PIMA dataset. The confusion matrix is employed to assess the classification's effectiveness. These matrix shows that the proposed model predicted 991 classes correctly in non-diabetic class. Similarly, 516 correctly predicted samples in diabetic class.

The classification algorithm's performance is summarized and visualized using the confusion matrix. Using a given set of test data, a confusion matrix is used to assess how well categorization models perform. It is calculated if the test data's real values are known.

Figure 6 shows the accuracy, f-measure, precision, TNR, and TPR. The experimental analysis indicates that the proposed approach outperforms the Random Forest, Naive Bayes, and Decision Tree models. The proposed method achieves the greatest accuracy of 98.43546%.

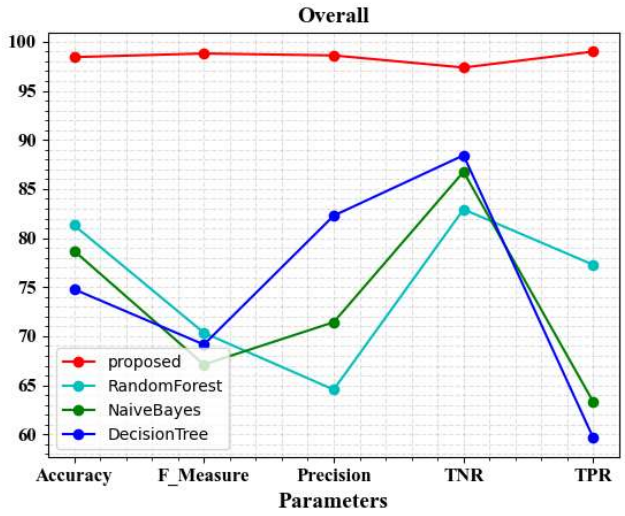

Figure-6 Comparison graph of accuracy, f-measure, precision, TNR and TPR

The Matthews Correlation Coefficient (MCC) is a robust evaluation metric for binary classifiers that assesses prediction quality by jointly accounting for true and false outcomes across both classes. By incorporating true positives, true negatives, false positives, and false negatives, MCC provides a comprehensive assessment of model performance. Unlike simple accuracy measures, it remains reliable even in situations where class distributions are significantly imbalanced.

It is widely used and recognized in biomedical research as a balanced metric that is excellent for classifications with imbalance concerns. Adoption of MCC in biomedical research underscores its effectiveness in handling class imbalances and providing a reliable assessment of classification performance.

Mathematically, MCC is defined as:

$$\text{MCC} = \frac{(TP*TN) - (FP*FN)}{\sqrt{(TP+FP)(TP+\quad)(TN+FP)(TN+FN)}} \qquad (5)$$

While MCC may produce a maximum score only when predictions are entirely correct, its balanced nature makes it invaluable for evaluating models in scenarios where data imbalances are prevalent and accurate predictions are crucial. Figure 7 compares the MCC to the previous method and the new method to determine which method yields the best results.

The classification error rate (CER) quantifies the fraction of incorrectly predicted samples out of the total dataset and is commonly used to assess classifier performance. This metric offers a straightforward indication of how frequently a model produces incorrect predictions. Although the proposed approaches outperform conventional machine learning models, it should be noted that their optimization process does not necessarily involve directly minimizing a loss function that explicitly corresponds to classification error rates.

CER is defined as:

$$\text{Classification Error} = \frac{error}{total} = \frac{FP+FN}{TP+TN+FP+F} \qquad (6)$$

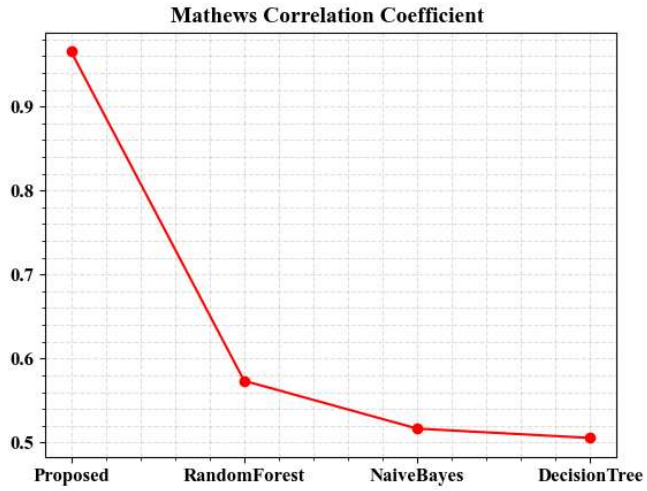

Figure-7 MCC comparison with the existing algorithm

Traditional machine learning algorithms often optimize loss functions such as cross-entropy, hinge loss, or squared error, which are not inherently tailored to minimizing the CER. Figure 8 compares CER with existing methods.

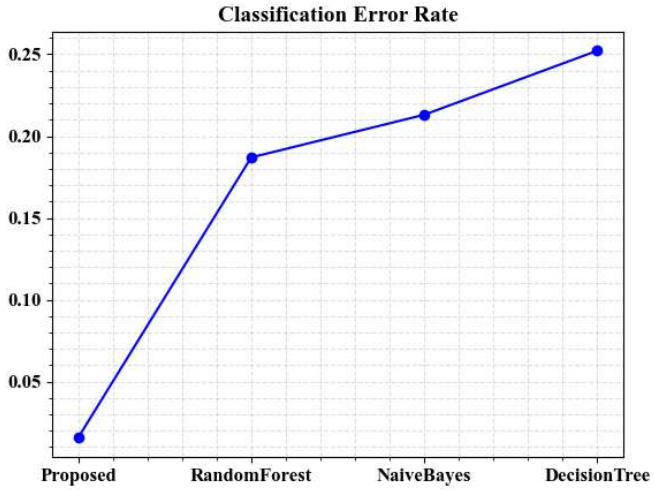

Figure-8 CER comparisons with existing methods

The evaluation incorporates both the false positive rate (FPR) and false negative rate (FNR) as key error indicators. A comparative analysis of these metrics with existing methods is presented in Figure 9. The proposed approach achieves an FPR of 0.0263 and an FNR of 0.263, demonstrating a notable reduction in errors. In contrast, the Random Forest model records FPR and FNR values of 0.1707 and 0.2273, respectively, while the Naive Bayes classifier yields values of 0.1325 for FPR and 0.3671 for FNR. The Decision Tree method reports corresponding values of 0.1157 and 0.4037. Overall, the proposed technique consistently exhibits lower error rates compared to the existing approaches.

FNR & FPR are defined as:

$$\text{FNR} = \frac{FN}{TP+F} = 1 - TPR \quad (7)$$

$$\text{FPR} = \frac{FP}{FP+TN} = 1 - TNR \quad (8)$$

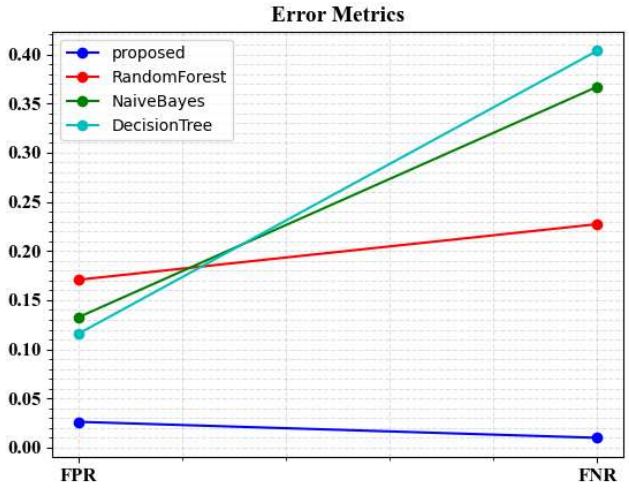

Figure-9 Error metrics comparison with existing methods

The loss value represents how well or poorly the model performs following iterative optimization. The accuracy measure is used to quantify an algorithm's performance in a human-readable manner.

Figure 10 shows accuracy vs. loss for the MW-FFT-OAconv classifier. The proposed method achieves superior accuracy while maintaining a comparatively low error rate.

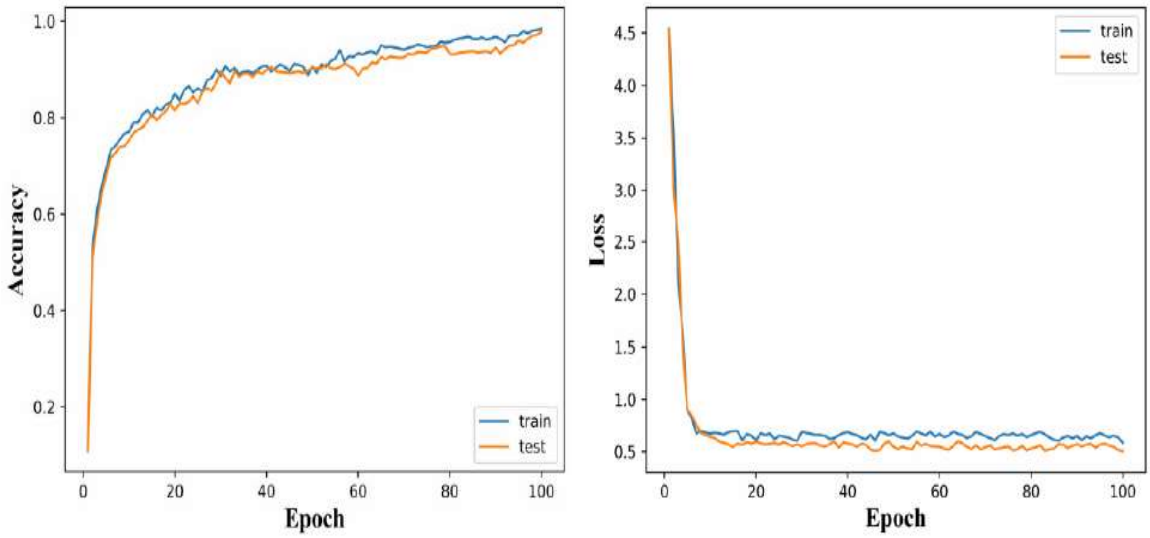

Figure-10 Accuracy vs. loss for MW-FFT-OAconv classifier

The error occurred during the process also measured as root mean square error and plotted in Figure 11. The proposed model possesses less error rate as compared to other models such as SVR, GPR and XGB (Tasin et al. [25]).

The RMSE value obtained by the proposed model is 0.22 while the SVR possess 0.45, XGB posses 0.36 and GPR possess 0.43. These values shows that the proposed model provides less error than other sue to its optimized parameter.

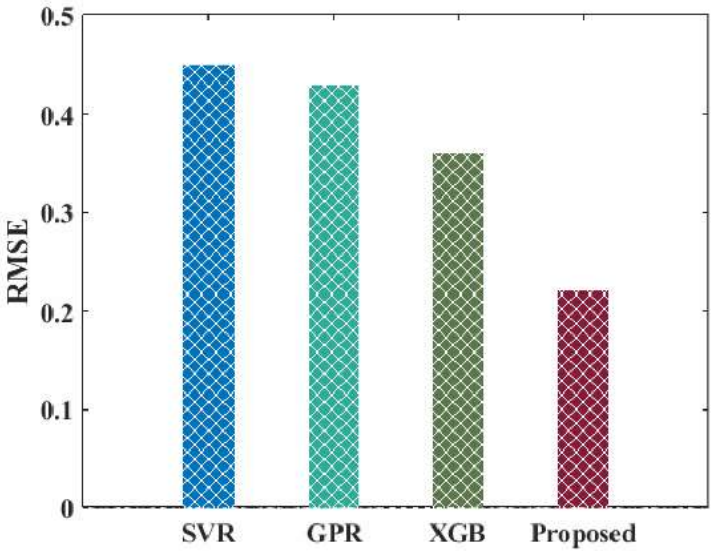

Figure-11 RMSE evaluation of proposed model

The optimized parameter is offer by the WMOV algorithm. Comparison table for various methods are presented in Table 2.

	Proposed	**Random Forest**	**NaiveBayes**	**Decision Tree**
Accuracy	98.43546	81.30435	78.69565	74.78261
Precision	98.60697	64.55696	71.42857	82.27848
F Measure	98.80359	70.34483	67.11409	69.14894
Specificity	97.37336	82.92683	86.75497	88.42975
Sensitivity	99.001	77.27273	63.29114	59.63303
MCC	0.965453	0.573436	0.516492	0.505374
CER	0.015645	0.186957	0.213043	0.252174
FPR	0.026266	0.170732	0.13245	0.115702
FNR	0.026266	0.227273	0.367089	0.40367

Table 2: Comparison Table for various methods

Table 3 presents the ablation analysis conducted on the proposed model.

Study	Accuracy Score (%)
Without Pre-processing	96.66
With Pre-processing	97.84
Without Data Augmentation	95.99
With Data Augmentation	96.53
Without Optimization	98.43
With Optimization	97.88

6.5 Implementation Setup

For the experiments, an Intel(R) Core(TM) i7-3610QM CPU running at 2.30GHz with four cores and four logical processors is employed. The computer's name is DESKTOP-FVQTL1J, Micro Software 10 pro, Microsoft Corporation is an operating system manufacturer, and it has 8GB of built-in RAM.

The developed model is implemented on the Python environment (version 3.9) and the experimental results is evaluated and compared with earlier machine learning prediction models with respect to statistical measures such as Accuracy, Precision, Classification Error Rate, F1-Score, and Matthews Correlation Coefficient.

7 Achievements with respect to Objectives

- Successfully developed Data Augmentation model using AVAE by combining the benefits of Variational Auto-Encoder for data generation and capability of the adversarial learning for the latent representation.
- Implemented a Morlet wavelet–based activation function in the hidden layer to enhance the classifier's ability to map augmented feature representations to various diabetes states.
- Designed an FFT-based overlap-and-add convolution framework to lower the computational load and execution time of the deep learning classifier, and further enhanced network accuracy by applying a metaheuristic optimization strategy for optimal hyperparameter selection.
- Evaluated and compared performance of proposed model compared with earlier machine learning prediction models with respect to statistical measures with respect to Accuracy, F1-Score, Precision, Classifier Error Rate, and Mathew Coefficient Correlation.

8 Conclusion

In this research, type 2 diabetes prediction is carried out using a proposed deep learning–based model applied to the widely used PIMA Indian Diabetes Dataset. Data pre-processing is treated as a critical phase, as the quality of input data directly influences the robustness and reliability of predictive performance. The primary objective of the proposed pre-processing strategy is the effective identification and elimination of outliers and the proper handling of missing or invalid values. To achieve this, a mathematical formulation based on the Interquartile Range (IQR) of each attribute is employed. This method facilitates the identification and elimination of abnormal samples that lie far outside the typical data distribution, which helps suppress noise and enhances the model's ability to generalize. In addition, incomplete records are handled by imputing missing values with the respective feature means, maintaining dataset integrity while retaining its overall statistical properties.

Following data preparation, the proposed model is evaluated using multiple statistical performance indicators to provide a comprehensive assessment of its predictive capability. The evaluation is carried out using multiple performance indicators, including accuracy, precision, F1-score, true positive rate (TPR), true negative rate (TNR), classification error percentage (CEP), and Matthews Correlation Coefficient (MCC). The results are then rigorously benchmarked against established machine learning models—such as Random Forest, Naïve Bayes, and Decision Tree classifiers—to validate the effectiveness and improved performance of the proposed framework.

By using the PIMA dataset, the presented deep learning–based predictive algorithm successfully estimates the probability of individual developing type 2 diabetes. Experimental results indicate that the proposed model outperforms conventional machine learning techniques across all evaluated metrics. Specifically, the model attains an accuracy of 98.44%, a precision of 98.61%, and an F1-measure of 98.80%, reflecting its strong classification capability. In terms of class-wise performance, the specificity (true negative rate) is recorded at 97.37%, while the sensitivity (true positive rate) reaches 99%, highlighting the effectiveness of the model in accurately classifying diabetic cases. Additionally, the Matthews correlation coefficient of 96% confirms a high level of prediction reliability, even in the presence of class imbalance. A low classification error rate (CER) of 0.015645, combined with minimal false positive (0.026266) and false negative (0.00999) rates, further highlights the robustness and clinical applicability of the proposed method.

Future research directions will focus on incorporating dietary and lifestyle-related parameters into the predictive framework, enabling a more personalized and holistic approach to type 2 diabetes management. By integrating individualized nutritional patterns with clinical data, the model can be extended to support proactive disease prevention and tailored treatment strategies, thereby enhancing its practical applicability in real-world healthcare settings.

9 Publications

Sr. No.	Title	Journal
1	Prediction of type 2 diabetes based on feature augmentation and Morlet wavelet assisted deep learning network with FFT overlap and add convolution	Published in International Journal of Biomedical Engineering and Technology (IJBET), Vol. 47, No. 3, 2025. DOI: 10.1504/IJBET.2025.144949
2	Data Augmentation approach for Type-2 Diabetes Prediction and Classification	Published in Indian Journal of Science and Technology 2025; 18(14):1096–1104. DOI: 10.17485/IJST/v18i14.3904

References

[1] S. Dey, B. Choudhury, and S. Sharanya, "Diabetes and its Complication Prediction using Multi-Task Learning," vol. 3075, no. 5, pp. 1426–1430, 2020, doi: 10.35940/ijitee.E2821.039520.

[2] B. Foretz, M., Guigas, B. & Viollet, "Understanding the glucoregulatory mechanisms of metformin in type 2 diabetes mellitus," *Nat. Rev. Endocrinol.*, vol. 15, pp. 569–589, 2019.

[3] G. Wang and B. Zhang, "Machine Learning For Tuning , Selection , And Ensemble Of Multiple Risk Scores For Predicting Type 2 Diabetes," pp. 189–198, 2019.

[4] A. Artasensi, A. Pedretti, G. Vistoli, and L. Fumagalli, "Type 2 diabetes mellitus: A review of multi-target drugs," *Molecules*, vol. 25, no. 8, pp. 1–20, 2020, doi: 10.3390/molecules25081987.

[5] N. Veronese, C. Cooper, and J. Reginster, "Europe PMC Funders Group Type 2 diabetes mellitus and osteoarthritis," vol. 49, no. 1, pp. 9–19, 2020, doi: 10.1016/j.semarthrit.2019.01.005.Type.

[6] M. Talebi Moghaddam *et al.*, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 220, 2024, doi: 10.1186/s12874-024-02341-z.

[7] M. M. Chowdhury, R. S. Ayon, and M. S. Hossain, "An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset," *Healthc. Anal.*, vol. 5, no. December 2023, p. 100297, 2024, doi: 10.1016/j.health.2023.100297.

[8] M. Agraz, Y. Deng, G. E. Karniadakis, and C. S. Mantzoros, "Enhancing severe hypoglycemia prediction in type 2 diabetes mellitus through multi-view co-training machine learning model for imbalanced dataset," *Sci. Rep.*, vol. 14, no. 1, p. 22741, 2024, doi: 10.1038/s41598-024-69844-z.

[9] S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Computer Methods and Programs in Biomedicine Update Improving SVM performance for type II diabetes prediction with an improved non-linear kernel : Insights from the PIMA dataset," *Comput. Methods Programs Biomed. Updat.*, vol. 4, no. August, p. 100118, 2023, doi: 10.1016/j.cmpbup.2023.100118.

[10] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, 2022, doi: 10.3390/s22197268.

[11] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-78519-8.

[12] R. F. Albadri, S. M. Awad, A. S. Hameed, T. H. Mandeel, and R. A. Jabbar, "A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm," *Math. Model. Eng. Probl.*, vol. 11, no. 8, pp. 2119–2126, 2024, doi: 10.18280/mmep.110813.

[13] 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," *Comput. Methods Programs Biomed.*, vol. 202, 2021, doi: 10.1016/j.cmpb.2021.105968.

[14] H. Shao, M. Xia, M. Ieee, J. Wan, and C. W. De Silva, "Modified Stacked Auto - encoder Using Adaptive Morlet Wavelet for Intelligent Fault Diagnosis of Rotating Machinery," vol. 4435, no. c, 2021, doi: 10.1109/TMECH.2021.3058061.

[15] K. Chitsaz, M. Hajabdollahi, N. Karimi, S. Samavi, and S. Shirani, "Acceleration of Convolutional Neural Network Using FFT-Based Split Convolutions," 2020, [Online]. Available: http://arxiv.org/abs/2003.12621

[16] I. Ahmadianfar, A. A. Heidari, S. Noshadian, H. Chen, and A. H. Gandomi, "INFO: An efficient optimization algorithm based on weighted mean of vectors," *Expert Syst. Appl.*, vol. 195, no. January, 2022, doi: 10.1016/j.eswa.2022.116516.

[17] B. M. K. P, S. P. R, R. K. Nadesh, and K. Arivuselvan, "International Journal of Cognitive Computing in Engineering Type 2 : Diabetes mellitus prediction using Deep Neural Networks classifier," *Int. J. Cogn. Comput. Eng.*, vol. 1, no. October, pp. 55–61, 2020, doi: 10.1016/j.ijcce.2020.10.002.

[18] V. Chang, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 0123456789, 2022, doi: 10.1007/s00521-022-07049-z.

[19] E. Babalola, T. Igbe, F. Yetunde, Z. Nie, and W. Lei, "Jou rna lP," *Futur. Gener. Comput. Syst.*, 2020, doi: 10.1016/j.future.2020.10.035.

[20] L. Al-Dabbasa and A. A. Abu-Sharehaa, "Early Detection of Female Type-2 Diabetes using Machine Learning and Oversampling Techniques," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1237–1245, 2024, doi: 10.47738/jads.v5i3.298.

[21] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Front. Genet.*, vol. 14, no. October, pp. 1–15, 2023, doi: 10.3389/fgene.2023.1252159.

[22] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation," *Adv. Human-Computer Interact.*, vol. 2022, 2022, doi: 10.1155/2022/9220560.

[23] F. J. Moreno-Barea, J. M. Jerez, and L. Franco, "Improving classification accuracy using data augmentation on small data sets," *Expert Syst. Appl.*, vol. 161, p. 113696, 2020, doi: 10.1016/j.eswa.2020.113696.

[24] E. Sabitha and M. Durgadevi, "Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 921–930, 2022, doi: 10.14569/IJACSA.2022.01309107.

[25] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.