Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

AN OPTIMIZED RESIDUAL CONVOLUTIONAL NEURAL NETWORK FRAMEWORK FOR EARLY DIABETES PREDICTION USING CLINICAL PARAMETERS

[1]Shams Ul Arifeen, [*2]Mian Farhan Shah, [3]Zia Ur Rahman, [4]Muhammad Naeem Ullah, [5]Waqas Ahmad, [6]Asad Khan, [7]Naeem Jan, [8]Atta Ur Rahman

[1]Department of Computer Science, Abdul Wali Khan University, Mardan
[*2]Department of Computer Science, Bacha Khan University, Charsadda
[3]Department of Computer Science, Bacha Khan University, Charsadda
[4]Department of Computer Science, Bacha Khan University, Charsadda
[5]Department of Computer Science, Bacha Khan University, Charsadda
[6]Department of Computer Science, Bacha Khan University, Charsadda
[7]Department of Computer Science, Bacha Khan University, Charsadda
[8]Department of Computer Science, Bacha Khan University, Charsadda
[*2]mianfarhanshah5@gmail.com

DOI:-https://doi.org/10.5281/zenodo.21170454

Keywords
Diabetes mellitus; Deep learning; Residual convolutional neural network; Machine learning; SMOTE; Prediction

Article History
Received: 23 May, 2026
Accepted: 24 June, 2026
Published: 26 June, 2026

Corresponding Author: *

Abstract
Diabetes mellitus is a chronic metabolic disorder with increasing global prevalence and serious health complications. Early prediction is important for prevention and effective management. This study developed an optimized Residual Convolutional Neural Network (RCNN) framework for diabetes prediction using clinical parameters from the Pima Indians Diabetes Dataset. Data preprocessing, normalization, and Synthetic Minority Over-sampling Technique (SMOTE) were applied to improve data quality and address class imbalance. The proposed RCNN model combines convolutional feature extraction with residual learning to identify complex nonlinear relationships among diabetes-related factors. Model performance was evaluated using accuracy, sensitivity, specificity, and Matthews Correlation Coefficient. The findings demonstrate that deep learning-based prediction can provide an effective computational approach for diabetes screening and clinical decision support. The proposed RCNN model achieved an accuracy of 96.41%, sensitivity of 96.17%, specificity of 97.53%, MCC of 0.94, and AUC of 0.99 on the testing dataset, outperforming conventional machine learning classifiers including Random Forest, Support Vector Machine, Decision Tree, and Extra Trees.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by persistent elevation of blood glucose levels due to impaired insulin secretion, reduced insulin sensitivity, or a combination of both mechanisms [1]. It represents one of the most challenging global health concerns because of its increasing prevalence and association with several life-threatening complications. Long-term uncontrolled diabetes can result in cardiovascular diseases, renal dysfunction, neuropathy, retinopathy, and other systemic complications that significantly affect the quality of life of patients. The continuous increase in diabetes prevalence has placed a substantial burden on healthcare systems worldwide, emphasizing the importance of early detection, effective monitoring, and timely therapeutic intervention [2].

The early identification of individuals at high risk of developing diabetes is essential because appropriate lifestyle modifications, preventive strategies, and medical interventions can delay or reduce disease progression. Conventional diagnostic approaches mainly depend on biochemical investigations, clinical evaluation, and interpretation of patient history. Although these methods are reliable, they may require extensive laboratory facilities, trained personnel, and repeated testing. Furthermore, the increasing availability of electronic health records and clinical datasets has created opportunities for developing computational approaches capable of analyzing large-scale medical information and assisting healthcare professionals in decision-making [3].

In recent years, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has gained significant attention in biomedical research due to its ability to discover hidden patterns from complex datasets. Machine learning algorithms have been extensively applied for the prediction and classification of various diseases, including diabetes mellitus. Several conventional ML techniques, such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree, Logistic Regression, and K-Nearest Neighbor, have been investigated for diabetes prediction using clinical parameters [4, 5]. These algorithms have demonstrated promising performance by identifying relationships between patient characteristics and disease outcomes [6].

However, conventional machine learning models often depend on manually selected features and predefined assumptions about the relationships between variables. Medical datasets usually contain complex nonlinear interactions among different physiological parameters, making feature extraction a challenging task. For example, factors such as glucose concentration, body mass index (BMI), age, insulin level, and genetic predisposition may interact with each other in complicated ways that are not easily recognized by traditional algorithms. Therefore, there is a need for advanced computational methods capable of automatically learning meaningful representations from medical data [7].

Deep learning approaches have emerged as powerful alternatives because they can automatically extract high-level features and identify complex patterns without extensive manual feature engineering [4]. Deep neural networks consist of multiple interconnected layers that transform input information into meaningful representations, enabling them to solve complex classification problems. In healthcare applications, deep learning models have demonstrated superior performance in disease prediction, medical imaging analysis, and clinical decision-support systems [8].

Among deep learning architectures, Convolutional Neural Networks (CNNs) have shown excellent capability in feature extraction and pattern recognition. Although CNNs were initially developed for image analysis, their ability to identify local and hierarchical feature relationships has encouraged their application in structured medical datasets [9]. However, increasing network depth may introduce optimization challenges, including vanishing gradients and degradation of learning performance [10].

To overcome these limitations, residual learning architectures were introduced. Residual Neural Networks (ResNet) utilize skip connections that allow information to bypass one or more layers, improving gradient flow and facilitating the training of deeper networks [11]. Residual connections help preserve important feature information and enhance model stability during optimization. The integration of convolutional operations with residual learning provides an effective strategy for extracting complex clinical features associated with disease development [12].

Another important challenge in medical classification problems is the imbalance between disease and non-disease samples. In many healthcare datasets, the number of healthy individuals is higher than patients with the target disease. Such imbalance can cause predictive models to become biased toward the majority class, resulting in poor identification of disease-positive cases. To address this problem, data balancing approaches such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely applied. SMOTE improves classification performance by generating synthetic samples for minority classes and providing a more balanced training environment [13].

In the present study, an optimized Residual Convolutional Neural Network (RCNN)-based framework integrated with SMOTE was developed for early prediction of diabetes mellitus using clinical parameters. The proposed approach combines preprocessing, data balancing, and advanced deep learning techniques to improve prediction reliability [14]. The developed framework aims to enhance the identification of diabetic individuals by learning complex relationships among important clinical attributes, including glucose level, BMI, insulin concentration, age, blood pressure, and diabetes-related risk factors. This computational approach may provide an effective decision-support tool for early diabetes screening and contribute toward the development of intelligent healthcare systems .

2. Materials and Methods

2.1 Dataset Description

In the present study, the Pima Indians Diabetes Dataset (PIDD) was utilized for the development and evaluation of the proposed diabetes prediction framework. This dataset is one of the most widely used benchmark datasets for evaluating computational models for diabetes mellitus classification. The dataset was originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases and contains clinical information of female patients of Pima Indian heritage with the objective of predicting the occurrence of diabetes [2, 15].

The dataset consists of multiple clinical and demographic attributes that are directly associated with diabetes risk factors. These variables provide important physiological information related to glucose metabolism, body composition, and genetic predisposition. The input features include the number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. The final outcome variable represents the class label, where individuals are categorized as diabetic or non-diabetic.

Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

The plasma glucose concentration is considered one of the most influential predictors because elevated glucose levels are strongly associated with impaired glucose regulation and diabetes development. Similarly, BMI provides information about obesity-related metabolic risk, whereas insulin level reflects abnormalities in insulin production and utilization. Age and diabetes pedigree function contribute additional information related to demographic and hereditary factors.

Table 1: *Description of clinical attributes used in the study*

S.No	Attributes	Description	Calculation	Range
1	Pregnancies	How many times pregnancy?	Years	0 to 3.0
2	Glucose	Plasma_glucose_level after 2 h	Mg/dl	0 to 199.0
3	Blood_Pressure	Individual_BP	(MM/Hg)	0 to 122.0
4	Skin_Tickness	Triceps fold thickness	MM	0 to 99.0
5	Insulin	Patient- blood_insulin-level	μ U/ml	0 to 846.0
6	BMI	Body-mass_index	Kg.	0 to 67,0
7	Diabetes Pedigree Function	Represents hereditary risk information associated with diabetes occurrence	1 = diabetic disease	0 to 2.45
8	Age	Age_of_individual	Years	1 to 78.0
9	Class	Class attribute	1 diabetic and 0 non-diabetic	0, 1

Before model development, the dataset was examined to understand feature distribution, variability, and class representation. Statistical analysis of the dataset parameters was performed to identify data characteristics and ensure appropriate preprocessing before model training.

Table 2: *Statistical characteristics of dataset variables*

S.No	Attributes	Counts	Mean Value	Std Deviation	Minimum	Maximum
1	Pregnancy	768	4.03	3.44	0	4.0
2	Glucose	768	0.55	0.49	0	1.0
3	Blood_Pressure	768	0.52	0.51	0	1.0
4	Skin-Tickness	768	0.5	0.49	0	1.0
5	Insulin	768	0.62	0.48	0	1.0
6	MBI	768	0.49	0.51	0	1.0
7	Diabetes_Pedigree_ Function	768	0.24	0.42	0	1.0
8	Age.	768	48.22	12.29	16	90.0

9	Class	768	0.52	0.480	0	1.0

2.2 Data Preprocessing and Class Balancing

Data preprocessing is an essential step in developing reliable machine learning and deep learning models because the quality of input data directly affects predictive performance. The clinical dataset was initially analyzed for inconsistencies, missing values, and abnormal observations. Since medical datasets often contain variations in numerical ranges among different attributes, normalization was performed to transform all features into a comparable scale [16].

Feature normalization reduces the dominance of variables with larger numerical values and allows the deep learning model to efficiently optimize during training. This step is particularly important for neural network-based approaches because large variations among input variables may negatively affect convergence and model stability [17].

Another important challenge associated with medical classification datasets is class imbalance. In the Pima Indians Diabetes Dataset, the distribution of diabetic and non-diabetic cases is not equal, resulting in a higher number of samples belonging to the majority class. If this imbalance is ignored, predictive models may become biased toward the majority class and show reduced ability to detect diabetic patients.

To overcome this limitation, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the training dataset [1]. SMOTE generates artificial samples for the minority class by analyzing neighboring samples and creating new synthetic observations between existing data points. This approach improves representation of diabetic cases and enables the model to learn meaningful patterns from both classes.

The balanced dataset was subsequently divided into training and testing subsets. The training data were used for model learning, whereas the testing data were reserved for independent evaluation of prediction performance.

2.3 Proposed Residual Convolutional Neural Network (RCNN) Model

A deep learning-based Residual Convolutional Neural Network (RCNN) framework was developed to improve diabetes prediction performance. The proposed architecture combines the feature extraction capability of convolutional neural networks with residual learning mechanisms to efficiently capture complex relationships among clinical variables.

The input clinical features were provided to the one-dimensional convolutional layer (Conv1D), which extracts important local patterns and interactions among diabetes-related attributes. The convolution operation allows the network to automatically identify significant feature combinations without requiring manual feature selection.

The convolutional layer was followed by a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity into the model and enables the network to learn complex relationships present within clinical data. A pooling layer was incorporated to reduce feature dimensionality and remove redundant information, thereby improving computational efficiency.

The major component of the proposed architecture is the residual block. Residual blocks contain skip connections that directly transfer information from earlier layers to deeper layers. These connections improve gradient propagation and reduce optimization difficulties commonly observed in deeper neural networks. The residual mechanism allows the network to retain important information while learning additional feature representations.

Batch normalization was applied to stabilize the learning process by reducing internal variation during training. This technique improves convergence speed and enhances model generalization. To prevent overfitting, dropout regularization was introduced by randomly disabling a proportion of neurons during training, forcing the model to learn more robust feature representations.

Finally, the extracted features were passed through fully connected dense layers followed by a sigmoid activation function. The sigmoid output layer produces a probability score representing the likelihood of diabetes occurrence. A threshold value was then applied to classify individuals into diabetic and non-diabetic groups. The model was trained using the Adam optimizer with a learning rate of 0.001 for 40 epochs with a batch size of 64. Dropout regularization (0.2) was applied to reduce overfitting.

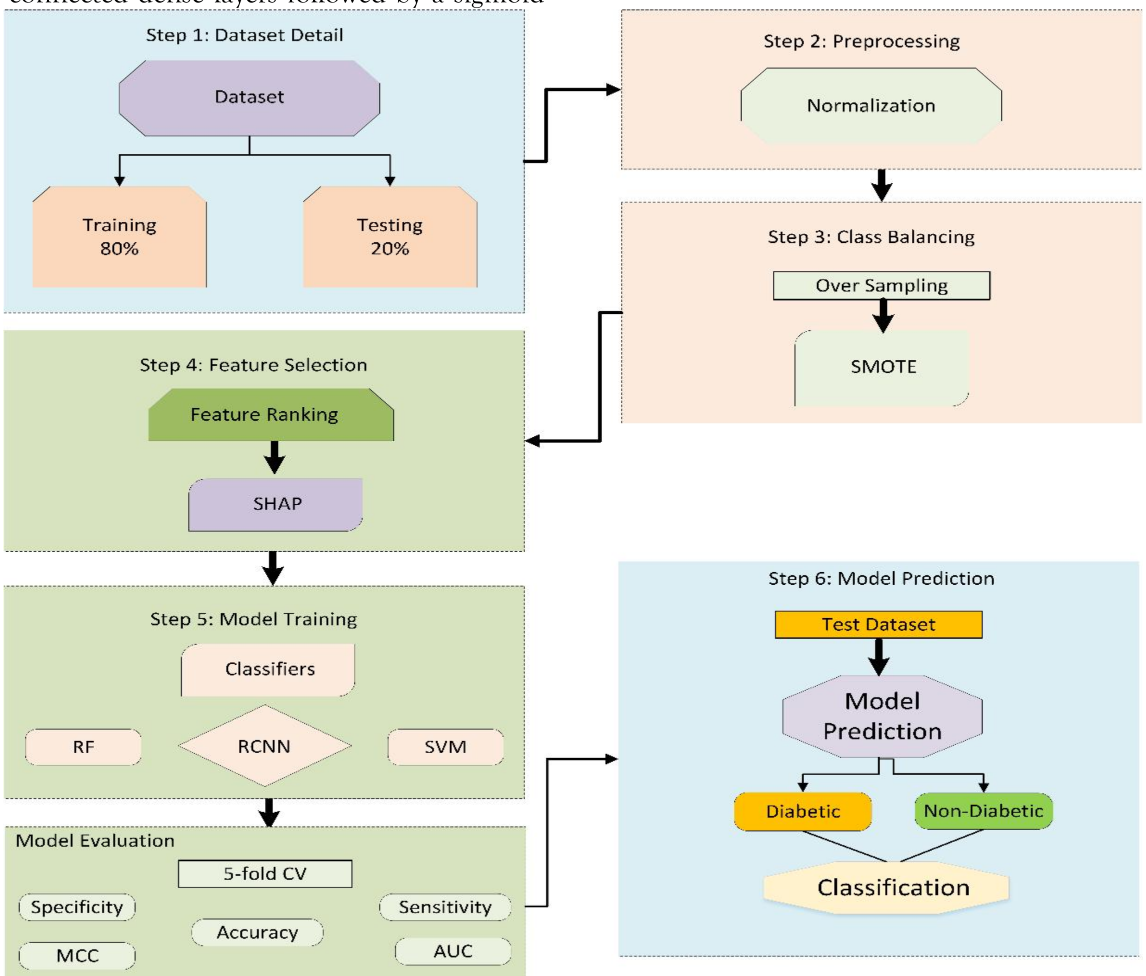


Figure 1. Proposed RCNN architecture for diabetes prediction

The complete workflow of the proposed framework consists of dataset collection, preprocessing, SMOTE-based balancing, RCNN model training, and performance evaluation using classification metrics including accuracy, sensitivity, specificity, and Matthews Correlation Coefficient. The model performance was evaluated using stratified 5-fold cross-validation to ensure reliable estimation and reduce sampling bias.

3. Results and Discussion

3.1 Dataset and Balancing Analysis

The distribution of diabetic and non-diabetic classes was analyzed before model training to evaluate the effect of class imbalance on the prediction framework. As shown in Figure 2, the original dataset exhibited an unequal distribution between the two classes, where non-diabetic samples represented the majority class while diabetic samples constituted the minority class. This imbalance may negatively influence the learning process because machine learning and deep learning models tend to become biased toward the dominant class, resulting in reduced sensitivity for detecting diabetic patients.

Before applying the balancing technique, the dataset contained a higher number of non-diabetic cases compared with diabetic cases [18]. Such an unequal distribution can lead to misleadingly high accuracy because the model may correctly classify a large proportion of majority-class samples while failing to recognize important minority-class diabetic cases. In clinical prediction problems, incorrect identification of diabetic individuals is a critical limitation because delayed diagnosis may prevent early intervention and disease management [19].

To overcome this limitation, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset. SMOTE generates new synthetic samples for the minority class by considering the similarity between existing minority observations and their nearest neighbors in the feature space. Unlike simple duplication methods, SMOTE creates new representative samples that improve data diversity and reduce the possibility of overfitting [20].

After applying SMOTE, the distribution between diabetic and non-diabetic classes became balanced, allowing the proposed RCNN model to receive an equivalent representation of both classes during training. This balanced learning environment improved the capability of the model to identify disease-related patterns and enhanced its ability to correctly classify diabetic individuals. The improved class distribution contributed to better sensitivity, specificity, and overall prediction performance of the proposed framework.

Therefore, SMOTE-based balancing played an important role in improving the reliability of the developed diabetes prediction model by minimizing class bias and supporting more accurate clinical classification.

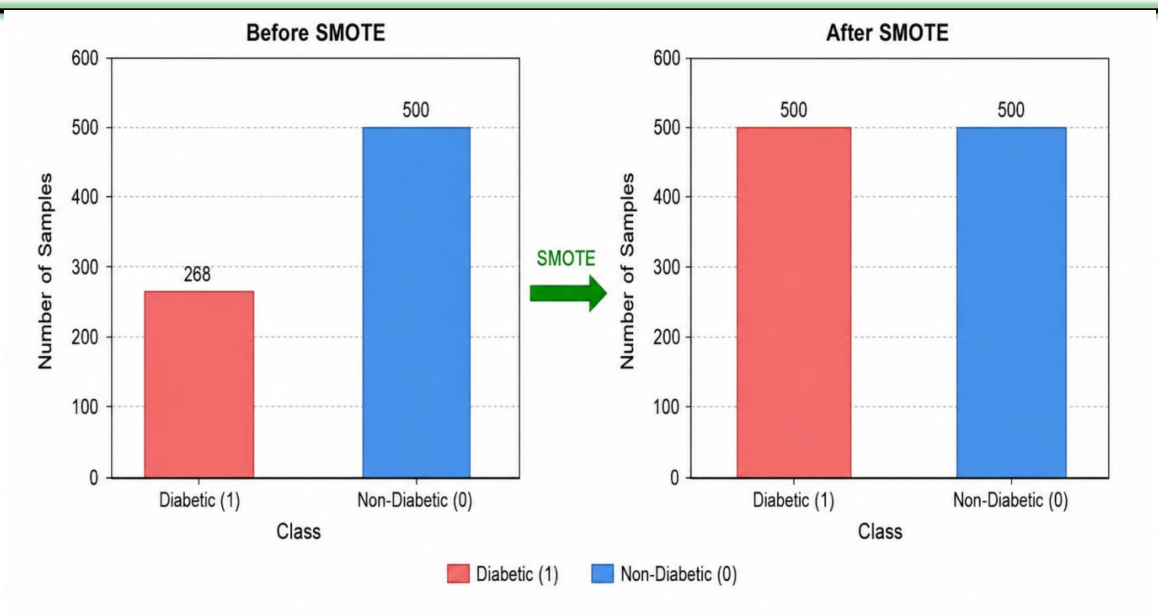


Figure 2. Class distribution before and after SMOTE-based data balancing.

3.2 RCNN Model Performance

The RCNN model demonstrated effective learning of complex associations among clinical parameters. Residual connections improved gradient propagation and allowed deeper feature learning.

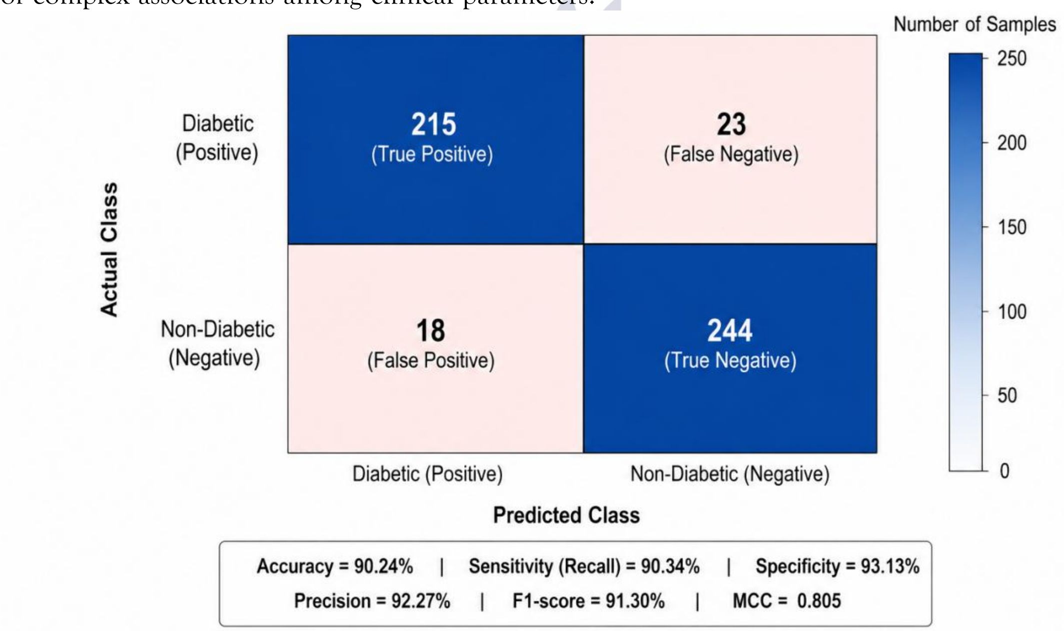


Figure 3. Confusion matrix of RCNN model.

The confusion matrix demonstrates the number of correctly and incorrectly classified diabetic and non-diabetic cases. The lower number of false-negative predictions indicates improved capability of the model in identifying diabetic patients.

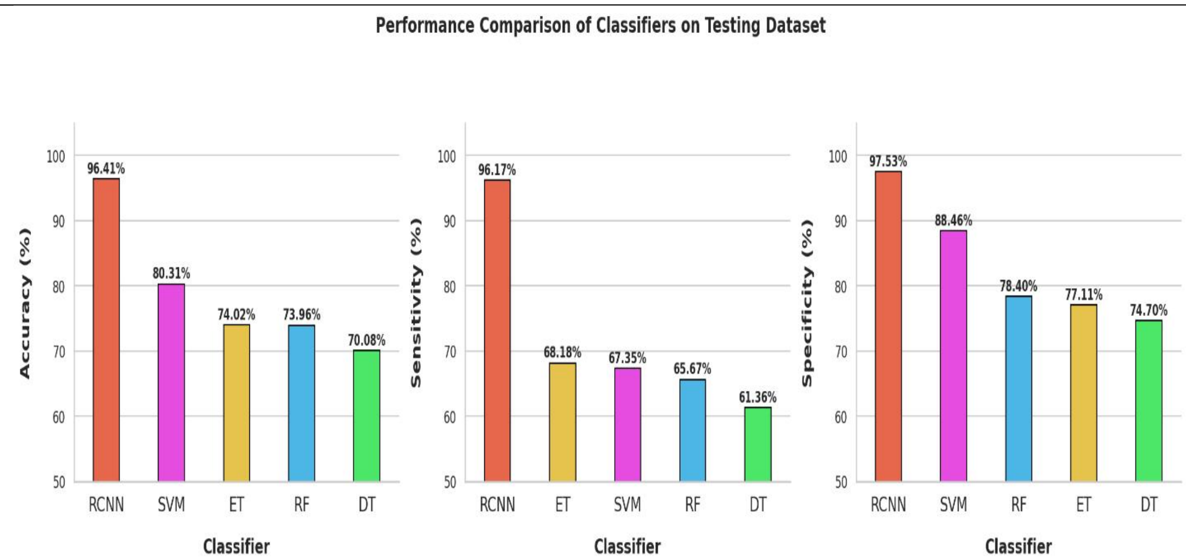
3.3 Comparison with Machine Learning Algorithms

Traditional algorithms such as Random Forest and Support Vector Machine provide useful classification ability but depend on predefined feature relationships. The RCNN model automatically extracts hierarchical features, resulting in improved predictive capability.

Table 3. Performance comparison of RCNN and traditional machine learning classifiers on testing dataset & Values represent testing dataset performance.

Include:

Model	Accuracy	Sensitivity	Specificity	MCC	AUC
RCNN	96.41	96.17	97.53	0.94	0.99
RF	73.96	65.67	78.40	0.44	0.81
SVM	80.31	67.35	88.46	0.58	0.86
DT	70.08	61.36	74.70	0.35	0.75
ET	74.02	68.18	77.11	0.44	0.74



3.4 ROC Analysis

Receiver Operating Characteristic (ROC) curve analysis was performed to evaluate the discrimination ability of the proposed RCNN model and compare its classification performance with conventional machine learning algorithms. The ROC curve represents the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds [21].

A model with higher sensitivity and lower false positive rate produces a curve closer to the upper-left corner, indicating better classification capability. The Area Under the Curve (AUC) was used as an overall measure of model performance, where a value closer to 1 represents excellent discrimination between diabetic and non-diabetic cases, while a value close to 0.5 indicates performance similar to random classification [22, 23].

The ROC analysis demonstrated that the proposed RCNN model achieved superior discriminative performance compared with traditional classifiers. The RCNN model obtained an AUC value of 0.99, indicating excellent ability to distinguish between diabetic and non-diabetic individuals. In comparison, conventional machine learning approaches including Random Forest, Support

Vector Machine, Decision Tree, and Extra Trees showed comparatively lower AUC values.

The improved ROC performance of the RCNN model can be attributed to its ability to automatically learn complex feature representations from clinical parameters. The combination of convolutional feature extraction and residual learning enabled the model to identify hidden relationships among diabetes-associated factors, resulting in improved sensitivity and specificity.

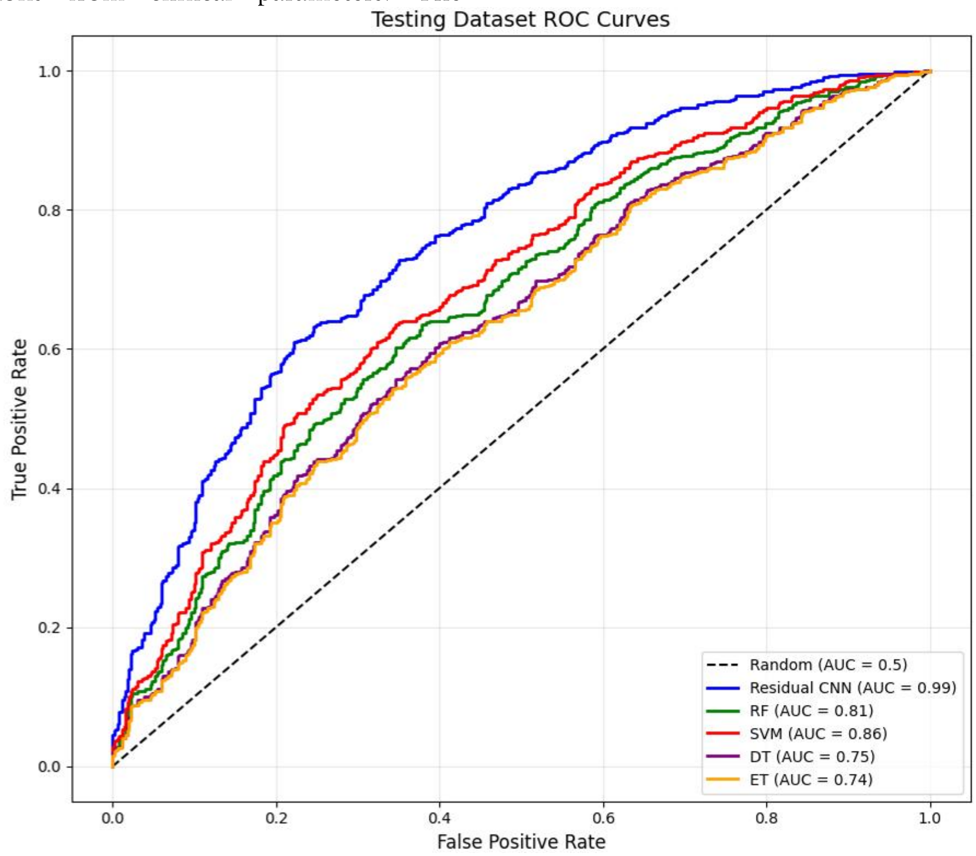


Figure 4. ROC curve of proposed RCNN model.

A higher area under the curve indicates better classification ability and stronger clinical applicability.

4. Conclusion

This study developed an optimized RCNN-based diabetes prediction framework using clinical parameters. Integration of SMOTE and residual learning improved model robustness and classification capability. The proposed approach can support early diabetes screening and future intelligent healthcare systems. Future studies may focus on validation using larger multi-center clinical datasets and integration with real-time healthcare decision-support systems.

Funding

No funding received from any source

References

1. Javed, M.A., et al., *Evaluation of pyrimidine/pyrrolidine-sertraline based hybrids as multitarget anti-Alzheimer agents: In-vitro, in-vivo, and computational studies.* Biomedicine & Pharmacotherapy, 2023. **159**: p. 114239.
2. Ejaz, I., et al., *Rational design, synthesis, antiproliferative activity against MCF-7, MDA-*

MB-231 cells, estrogen receptors binding affinity, and computational study of indenopyrimidine-2, 5-dione analogs for the treatment of breast cancer. Bioorganic & Medicinal Chemistry Letters, 2022. **64**: p. 128668.

3. Mahnashi, M.H., et al., *Neuroprotective potentials of selected natural edible oils using enzyme inhibitory, kinetic and simulation approaches.* BMC complementary medicine and therapies, 2021. **21**(1): p. 248.
4. Alam, W., et al., *In vitro 5-LOX inhibitory and antioxidant potential of isoxazole derivatives.* Plos one, 2024. **19**(10): p. e0297398.
5. Hassani, H., et al., *Artificial intelligence (AI) or intelligence augmentation (IA): what is the future?* Ai, 2020. **1**(2): p. 8.
6. Jan, M.S., et al., *Synthesis of pyrrolidine-2, 5-dione based anti-inflammatory drug: in vitro COX-2, 5-LOX inhibition and in vivo anti-inflammatory studies.* Latin Am J Pharm, 2019. **38**(11): p. 2287-2294.
7. Mahmood, F., et al., *Ethyl 3-oxo-2-(2, 5-dioxopyrrolidin-3-yl) butanoate derivatives: anthelmintic and cytotoxic potentials, antimicrobial, and docking studies.* Frontiers in chemistry, 2017. **5**: p. 119.
8. Alshehri, O.M., et al., *Succinimide Derivatives as Antioxidant Anticholinesterases, Anti-α-Amylase, and Anti-α-Glucosidase: In Vitro and In Silico Approaches.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 6726438.
9. Zhao, X., et al., *A review of convolutional neural networks in computer vision.* Artificial Intelligence Review, 2024. **57**(4): p. 99.
10. Waheed, B., et al., *Synthesis, antioxidant, and antidiabetic activities of ketone derivatives of succinimide.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 1445604.
11. Alshehri, O.M., et al., *Investigation of anti-nociceptive, anti-inflammatory potential and ADMET studies of pure compounds isolated from Isodon rugosus Wall. ex Benth.* Frontiers in pharmacology, 2024. **15**: p. 1328128.
12. Xu, G., et al., *Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey.* arXiv preprint arXiv:2405.01725, 2024.
13. Mahnashi, M.H., et al., *GC-MS Analysis and Various In Vitro and In Vivo Pharmacological Potential of Habenaria plantaginea Lindl.* Evidence-Based Complementary and Alternative Medicine, 2022. **2022**(1): p. 7921408.
14. Zafar, R., et al., *Prospective application of two new pyridine-based Zinc (II) amide carboxylate in management of alzheimer's disease: Synthesis, characterization, computational and in vitro approaches.* Drug Design, Development and Therapy, 2021: p. 2679-2694.
15. Ashour, A.F., et al. *Optimized neural networks for diabetes classification using pima Indians diabetes database.* in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI).* 2024. IEEE.
16. Masood, N., et al., *Antioxidant, carbonic anhydrase inhibition and diuretic activity of Leptadenia pyrotechnica Forssk. Decne.* Heliyon, 2023. **9**(12).
17. Shahzadi, K., et al., *Novel coumarin derivatives as potential urease inhibitors for kidney stone prevention and antiulcer therapy: from synthesis to in vivo evaluation.* Pharmaceuticals, 2023. **16**(11): p. 1552.
18. Pamungkas, B.P., M.J. Vikri, and I.A. Sa'ida, *Application of SMOTE-ENN Method in Data*

Balancing for Classification of Diabetes Health Indicators with C4. 5 Algorithm. Jurnal Sisfokom (Sistem Informasi dan Komputer), 2025. **14**(2): p. 183-188.

19. Ullah, K., et al., *Investigation of pivalic acid-derived organotin (IV) carboxylates: Synthesis, structural insights, interaction with biomolecules, and computational studies.* Journal of Molecular Structure, 2025. **1322**: p. 140444.
20. Rauf, A., et al., *Hypoglycemic, anti‐inflammatory, and neuroprotective potentials of crude methanolic extract from Acacia nilotica L.–results of an in vitro study.* Food Science & Nutrition, 2024. **12**(5): p. 3483-3491.
21. Ni, M., et al., *A deep learning approach for MRI in the diagnosis of labral injuries of the hip joint.* Journal of Magnetic Resonance Imaging, 2022. **56**(2): p. 625-634.
22. Pervaiz, A., et al., *Comparative in-vitro anti-inflammatory, anticholinesterase and antidiabetic evaluation: computational and kinetic assessment of succinimides cyano-acetate derivatives.* Journal of Biomolecular Structure and Dynamics, 2022: p. 1-14.
23. Nur-A-Alam, M., et al., *A faster RCNN-based diabetic retinopathy detection method using fused features from retina images.* IEEE Access, 2023. **11**: p. 124331-124349.