A Deep Learning-based Architecture for Diabetes Detection, Prediction, and Classification

Muhammad Hanfia Fakhar
Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, Pakistan
hanfiafakhar11@gmail.com

Muhammad Zeeshan Baig
Department of Information Technology, Wentworth Institute of Higher Education, Sydney, Australia
muhammad.baig@win.edu.au

Arshad Ali
Faculty of Computer and Information Systems, Islamic University of Madinah, Saudi Arabia
a.ali@iu.edu.sa

Muhammad Tausif Afzal Rana
School of Information Technology, King's Own Institute, Sydney, Australia
touseef.rana@koi.edu.au

Hamayun Khan
Department of Computer Science, Faculty of Computer Science & IT, Superior University Lahore, Pakistan
hamayun.khan@superior.edu.pk (corresponding author)

Waseem Afzal
School of Information Technology, VIC - Level 1, West Melbourne, Australia
waseem@pia.edu.au

Hafiz Umar Farooq
Department of Computer Science, Faculty of Computer Science and IT, Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan
umarfarooq4702@gmail.com

Sami Albouq
Faculty of Computer and Information Systems, Islamic University of Madinah, Saudi Arabia
salbouq1@iu.edu.sa

Received: 29 July 2024 | Revised: 18 August 2024 and 23 August 2024 | Accepted: 3 September 2024

 DOI: https://doi.org/10.48084/etasr.8354

ABSTRACT

This study examines the importance of Deep Learning (DL) in the Internet of Medical Things (IoMT) in providing impactful results in the diagnosis, classification, prediction, and categorization of stages of diabetes. A DL model was used to classify diabetic retinopathy data, based on a Multi-Layer Feed-Forward Neural Network (MLFNN). The Pima Diabetes Dataset (PDD) was used to train and test the proposed model. To increase accuracy, this study considered different activation functions and strategies

to deal with lost information. The proposed Multilayer Feed-Forward Neural Network (MLFNN) model was compared with conventional Machine Learning (ML) approaches, specifically Random Forest (RF) and Naive Bayes (NB), outperforming them with a significant increase in classification accuracy.

Keywords-Artificial Intelligence (AI); neural networks; Multilayer Feed-Forward Neural Network (MLFNN); Naive Bayes (NB); Random Forest (RF); Pima Diabetes Dataset (PDD)

I. INTRODUCTION

Finding medical anomalies is often seen as a complex undertaking that requires the knowledge of medical professionals. Expanded blood glucose levels have the potential to adversely affect essential organs, including the kidneys and the heart. Diagnosis of diabetes is one of the most complicated disarrangements that regularly require the collaboration of several doctors with different specialties [1]. Increased blood glucose levels have the potential to adversely affect essential organs, including the kidneys and the heart. Deciding whether an individual has a disorder can be a complex and challenging task that requires particular knowledge and abilities. The huge amount of data, in combination with the small sample sizes, emphasizes how DL strategies can accurately classify and recognize diabetic issues [2]. In [3], a strategy was developed based on a Multi-Layer Feed-Forward Neural Network (MLFNN) to detect and classify diabetes. The PDD dataset has missing values, which requires examining new techniques to address this problem. DL can not only optimize diagnostic issues but also classify medical issues that normally require labor-intensive strategies, providing reliable and effective results in less time [4]. This study explores DL for the classification of diabetes using PDD [5]. An input layer, one or more hidden layers, and an output layer comprise the MLFNN, which is a complete neural network design. A weight coefficient characterizes the association between two neurons [6, 7]. In [8], the foundation of the backpropagation strategy was investigated. Backpropagation works using two fundamental stages: the forward proliferation stage and the weight alteration stage. The layers of a neural arrangement are initialized with irregular weight values [9], and the output of each neuron is controlled by its activation function [10, 11]. Each layer in DL networks employs characteristics extracted from the previous layer in a progressive extraction procedure [12]. This multilevel strategy recombines characteristics from previous levels to enable more deep layers to distinguish perplexing details inside the data.

II. LITERATURE REVIEW

DL algorithms can handle massive datasets with viability, which is much appreciated for this characteristic. As many datasets are unstructured or lack labeling, DL algorithms can identify patterns and structures in datasets that may lack clear categorization [7]. Activation functions are essential elements in the architecture of neural networks. The output of a neuron is restricted by an activation function, which also controls the firing state of its connections [13-15]. Two distinct forms of activation functions are the Scaled Exponential Linear Unit (SELU) and the Exponential Linear Unit (ELU).

ELU with $0 < \alpha$ is given by:

$$f(x) = \begin{cases} \alpha(\exp(x) - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (1)$$

The hyperparameter in ELU controls the saturation level of negative inputs. In deep neural networks, ELU speeds up learning more quickly than other activation functions. Unlike other activation functions, ELU mitigates the vanishing gradient problem by incorporating negative values, which facilitate the adjustment of mean activation toward zero, a feature also observed in batch normalization. ELU achieves this with reduced computational overhead [16].

SELU is defined as:

$$f(x) = \lambda \begin{cases} \alpha(\exp(x) - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (2)$$

SELU allows for generating mappings consistent with the properties of Self-Normalizing Neural Networks (SNN). Neuron activations in SNNs tend to converge to zero mean and unit variance, as shown in (2). This characteristic enables reliable propagation across multiple layers, even in the face of noise and disturbances. As a result, SNNs make it easier to train deep neural networks with several layers by leveraging this effective regularization strategy, which improves learning robustness [17].

In [18], two classifiers were used, Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM), which utilized 13 components to analyze diabetic issues. MLP showed a 98% accuracy, while SVM achieved a slightly lower accuracy of 96%. In [19], diabetes was classified using a General Regression Neural Network (GRNN) based on nonparametric regression. In training, this model achieved an accuracy of 82.99%, while in testing, it achieved 80.21%. In [20], an open-source dataset was introduced and Naive Bayes, RBF, and J48 were used for diabetes prediction, achieving accuracies of 76.95%, 74.34%, and 76.5%, respectively.

III. METHODOLOGY

The experiments were carried out using the proposed model by dividing the PIMA dataset [21] into 70% for training and 30% for testing. The average performance over these iterations reliably assessed the model's stability and efficacy. The main steps included data collection, preprocessing, model selection, and evaluation. Figure 1 elaborates the flow chart of the proposed MLFNN architecture that effectively balances depth and width by enabling the model to capture complex patterns in the dataset and achieve high classification accuracy.

A. Data Collection

This study used the PIMA Diabetes Dataset (PDD) [21]. PDD is a free open-source dataset under a Universal Common license. Five pertinent features from the dataset, including 268 cases of diabetes and 500 cases of non-diabetes, were considered. The dataset includes standard values, as shown in Table I.

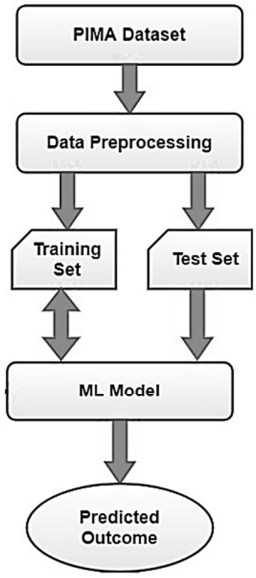

Fig. 1. Structure of the ML model

TABLE I. TEST FEATURES BASED ON THE PID DATASET

	Feature	Metric
1	High blood pressure diastolic	mmHg
2	Triceps skin folds	mm
3	Two-hour insulin level	U/ml
4	Index of muscle mass	kg/m^2
5	Diabetes pedigree function	U/ml

B. Proposed Model

The proposed MLFNN model for diabetes classification involves the following architecture and training process:

- Input layer: 10 neurons for the eight features plus two additional neurons for better handling nonlinearity.
- Hidden layers: Three hidden layers with 60 neurons each, employing ELU and SELU activation functions.
- Output Layer: Single neuron for binary classification.

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j \quad (3)$$

- Training algorithm: Backpropagation with forward and backward propagation phases.
- Evaluation: Monte Carlo cross-validation over 200 iterations.

Figure 2 illustrates the detection and classification model based on n samples by considering the feature matrix (X) to predict both diabetic and non-diabetic cases. The MLFNN architecture effectively balances depth and width, enabling the model to capture complex patterns in the dataset and achieve high classification accuracy.

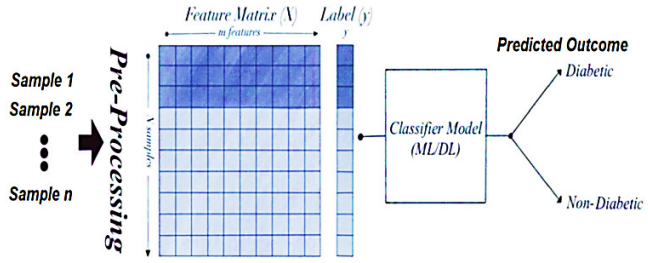

Fig. 2. Diabetes detection using the ML/DL models.

C. Activation Functions

Activation functions introduce non-linearity to the model. Two activation functions were examined. ELU helps accelerate the learning process by mitigating the vanishing gradient problem and promoting a zero-mean output. SELU supports self-normalizing neural networks by ensuring that the mean and variance of neuron activation approach zero.

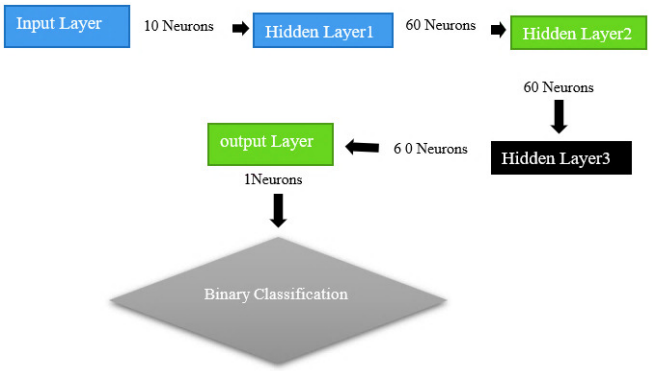

Fig. 3. Structure of proposed model.

D. Training Procedure

The MLFNN model was trained using the backpropagation algorithm, which operates in two phases. In forward propagation, initial weights are assigned randomly, and the network output is calculated.

$$a^{(l)} = \sigma\left(z^{(l)}\right) \quad (4)$$

In backward propagation, the calculated results are compared using a loss function to calculate the error, which is then propagated backward to adjust the weights iteratively until the error rate reaches an acceptable level.

$$\delta^{(l)} = \frac{\partial L}{\partial z^{(l)}} = \delta^{(l+1)} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial z^{(l)}} \quad (5)$$

E. Evaluation Metrics

The following metrics were used to evaluate the performance of the MLFNN model:

- Accuracy: Correctly predicted instances compared to the total instances.
- Recall: Correctly predicted positives to actual positives.
- Precision: Correctly predicted positives to total predicted positives.

F. Design

Figure 3 demonstrates the proposed design for diabetes diagnosis. The network's depth refers to the number of hidden layers, while its width signifies the total number of neurons across these layers. Although a single-layer feed-forward neural network can theoretically execute various learning tasks, it often suffers from limited learning rates and accuracy. In contrast, deeper learning models can reduce the number of neurons per layer, thus minimizing generalization errors.

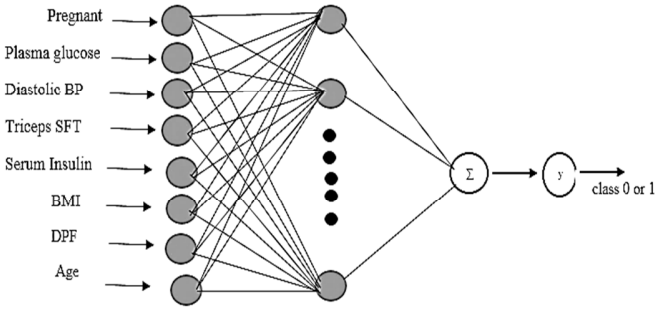

Fig. 4. Proposed MLFNN architecture for diabetes classification.

The Monte Carlo cross-validation approach was used to evaluate the MLFNN model, which is a periodic holdout validation process. 90% of the PDD dataset was split into training and 10% into validation sets. This was carried out iteratively 200 times using different seeds and the average performance was calculated. Employing randomly chosen test subsets, this approach ensures reliable evaluation, providing insight into its effectiveness and execution changeability over random test data tests.

G. Data Preprocessing

Applying specific preprocessing techniques to the input before integrating them into the system can enhance a neural network's training efficiency. The initial step involved examining potential correlations among the features. Correlation measures the extent to which one influences another input feature. Eliminating correlated features can accelerate the learning rate and mitigate biases within the neural network. In particular, correlation bias can adversely affect the performance of the Random Forest model.

Normalization is another significant preprocessing step to prepare the dataset. This method scales each input to a reliable extent, minimizing neural predisposition. Standardizing the data assists the preparation stage by ensuring that all features are inside comparable scales. This approach is especially advantageous when critical aberrations exist between inputs, including scales. The MinMaxScaler was used to normalize the PID dataset, which adapts all features to a predefined extent.

$$MinMaxScaler(feature\ range = (0,1);\ copy = True) \quad (6)$$

$$X_{std} = \frac{X - Xmin}{Xmax - Xmin} \quad (7)$$

$$X_{scaled} = X.std * max - min + min \quad (8)$$

The PDD dataset contains various lost values that can influence the classification performance of the neural network. Three different methods were examined to address this issue. First, all columns that contained lost values were removed. Then, all lost values were substituted with zeros. Third, lost values were replaced with the mean of the other values in the same column.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the proposed MLFNN was evaluated against Naïve Bayes (NB) and Random Forrest (RF). In addition, the results were compared for the different strategies for handling lost values.

A. Comparative Investigation

Table II shows the performance of NB and RF for diabetes classification. NB, known for its simplicity and robustness in classification tasks, yielded 75.22% and 71.55% training and testing accuracy, respectively. RF provided better results, as it can better handle missing values and categorical data. Table III highlights the superiority of the MLFNN model over NB and RF. Figure 5 the training and validation losses for the proposed model using ELU, which stabilizes after 100 epochs and remains moderately reliable after 500 epochs.

TABLE II. COMPARISON OF CLASSIFICATION ALGORITHMS

Data	NB	RF
Accuracy % (training)	75.22	98.1
Accuracy % (testing)	71.55	72.2

TABLE III. PROPOSED MODEL'S PERFORMANCE

	TA % (ELU)	TA % (SELU)
MLFNN TA (%)	92.11	86.55

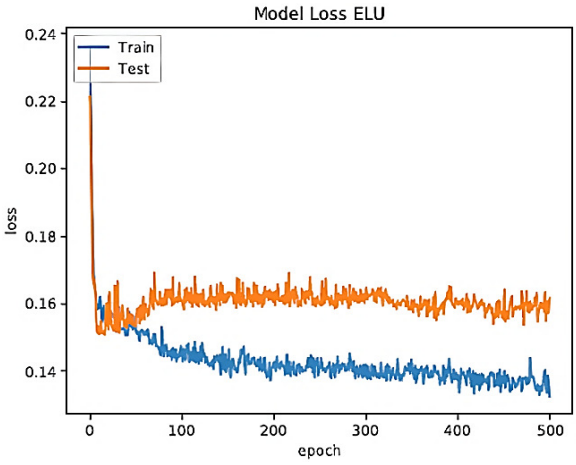

Fig. 5. Change in model loss in training and testing using ELU.

Figure 6 shows the training and validation losses of the proposed model using SELU as the activation function.

Figure 7 shows the Diped Function distribution in the PIMA dataset.

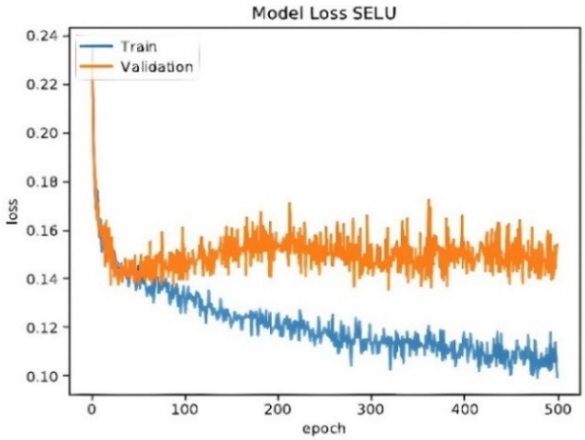

Fig. 6. Model loss in training and testing using SELU.

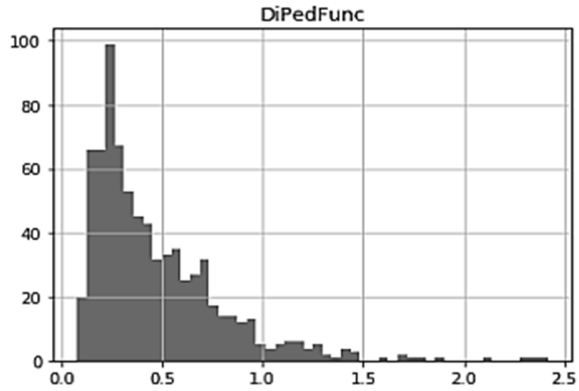

Fig. 7. Attributes of Diped Function.

Figure 8 shows the Skin Thick distribution. Due to high diabetes, the skin thickness and swelling ratio increases. The peak is above 200 u/ml over time, while using antidiabetics, it gradually decreases and drops below 50 u/ml.

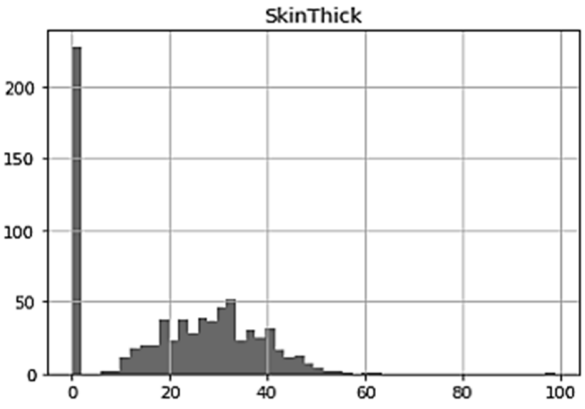

Fig. 8. Attributes of SkinThick.

Figure 9 shows the BloodP distribution. Increased blood sugar increases blood pressure levels. Tratio is getting high, having a peak above 80 u/ml with time, while using antidiabetics, it gradually decreases and drops below 10 u/ml.

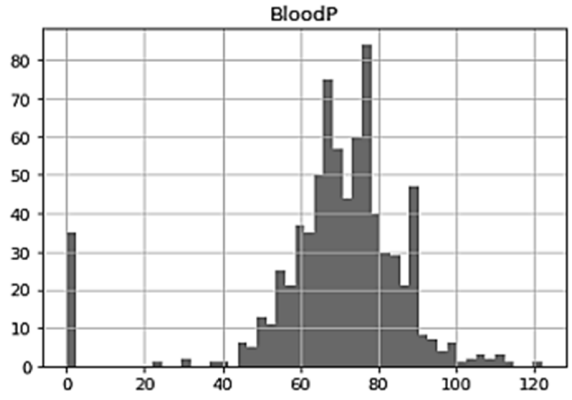

Fig. 9. Attributes of BloodP.

V. CONCLUSION

This study implemented a prediction and classification model for diabetes using a DL approach. The proposed MLFNN method outperformed the NB and RF classifiers in predicting diabetes using the PIMA Diabetes dataset. The best results for the proposed method were achieved by ignoring columns with lost values, instead of replacing lost values with zeros or means. The impact of ELU as an activation function is critical for improving neural network performance. After comparing different high-performance activation functions, it was shown that ELU was the best fit for the PDD dataset. Future studies should focus on developing an automated system in the form of a mobile application based on the proposed DL algorithm to help healthcare professionals in the early detection of diabetes.

REFERENCES

[1] R. Ramesh and S. Sathiamoorthy, "A Deep Learning Grading Classification of Diabetic Retinopathy on Retinal Fundus Images with Bio-inspired Optimization," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11248–11252, Aug. 2023, https://doi.org/10.48084/etasr.6033.

[2] C. Nithyeswari and G. Karthikeyan, "An Effective Heuristic Optimizer with Deep Learning-assisted Diabetic Retinopathy Diagnosis on Retinal Fundus Images," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14308–14312, Jun. 2024, https://doi.org/10.48084/etasr.7004.

[3] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, Dec. 2018, https://doi.org/10.1016/j.icte.2018.10.005.

[4] S. Gujral, "Early Diabetes Detection using Machine Learning: A Review," *International Journal for Innovative Research in Science & Technology*, vol. 3, no. 10, pp. 57–62, 2016.

[5] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, Dec. 2021, Art. no. 30, https://doi.org/10.1186/s42492-021-00097-7.

[6] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, Jan. 2019, https://doi.org/10.1016/j.procs.2020.01.047.

[7] B. Farajollahi *et al.*, "Diabetes diagnosis using machine learning," *Frontiers in Health Informatics*, vol. 10, Mar. 2021, Art. no. 65.

[8] R. Katarya and S. Jain, "Comparison of Different Machine Learning Models for diabetes detection," in *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*, Coimbatore, India, Dec. 2020, pp. 1–5, https://doi.org/10.1109/ICADEE51157.2020.9368899.

[9] M. Warke, V. Kumar, S. Tarale, P. Galgat, and D. J. Chaudhari, "Diabetes Diagnosis using Machine Learning Algorithms," *International Research Journal of Engineering and Technology*, vol. 06, no. 03, pp. 1470–1476, 2019.

[10] H. Khan, S. Ahmad, N. Saleem, M. U. Hashmi, and Q. Bashir, "Scheduling Based Dynamic Power Management Technique for offline Optimization of Energy in Multi Core Processors," *International Journal of Scientific & Engineering Research*, vol. 9, no. 12, pp. 6–10, 2018.

[11] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep Learning for Diabetes: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, Jul. 2021, https://doi.org/10.1109/JBHI.2020.3040225.

[12] A. H. Syed and T. Khan, "Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study," *IEEE Access*, vol. 8, pp. 199539–199561, 2020, fhttps://doi.org/10.1109/ACCESS.2020.3035026.

[13] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *International Journal of Information Technology*, vol. 12, no. 2, pp. 419–428, Jun. 2020, https://doi.org/10.1007/s41870-018-0270-5.

[14] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, Oct. 2019, Art. no. 101, https://doi.org/10.1186/s12902-019-0436-6.

[15] A. S. Alenizi and K. A. Al-karawi, "Machine Learning Approach for Diabetes Prediction," in *Proceedings of Eighth International Congress on Information and Communication Technology*, London, UK, 2024, pp. 745–756, https://doi.org/10.1007/978-981-99-3043-2_61.

[16] G. Bhuvaneswari and G. Manikandan, "A novel machine learning framework for diagnosing the type 2 diabetics using temporal fuzzy ant miner decision tree classifier with temporal weighted genetic algorithm," *Computing*, vol. 100, no. 8, pp. 759–772, Aug. 2018, https://doi.org/10.1007/s00607-018-0599-4.

[17] L. Math and R. Fatima, "Adaptive machine learning classification for diabetic retinopathy," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5173–5186, Feb. 2021, https://doi.org/10.1007/s11042-020-09793-7.

[18] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, Dec. 2021, Art. no. 30, https://doi.org/10.1186/s42492-021-00097-7.

[19] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *Journal of Healthcare Engineering*, vol. 2021, no. 1, 2021, Art. no. 9930985, https://doi.org/10.1155/2021/9930985.

[20] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, 2022, Art. no. 2789760, https://doi.org/10.1155/2022/2789760.

[21] V. Mishra, "Diabetes Prediction." Kaggle, [Online]. Available: https://kaggle.com/code/mvanshika/diabetes-prediction.