2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)

# AI/ML-Based Diabetes Application using Hybrid Grey Wolf and Dipper Throated Optimization Algorithm

1st Rahul Dattangire
*Independent Researcher*
USA
Rahuldattangire@ieee.org

2nd Saigurudatta Pamulaparthyvenkata
*Senior Data Engineer*
*Independent researcher*
USA
saigurudatta.pamulaparthyv@gmail.com

3rd Shreekant mandvikar
*Independent researcher*
USA
shreekant.mandvikar@gmail.com

4th Anandaganesh Balakrishnan
*Principal Software Engineer*
*Independent researcher*
USA
Anandaganesh.balakrishnan@gmail.com

5th Pradeep Chintale
*Independent Researcher*
*USA*
chintale.pradeep@gmail.com

2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) | 979-8-3503-6066-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/IACIS61494.2024.10721678

*Abstract*—**The application of Deep Learning (DL) in diagnosing chronic epidemiological disorders such as diabetes mellitus has become crucial due to the widespread occurrence of this disease worldwide. However, the diagnosis of diabetes faces many challenges such as missing values, high dimensionality of features, and accuracy issues. This paper proposes a Grey Wolf and Dipper Throated Optimization (GWDTO) algorithm for feature selection to address these challenges. The GWDTO algorithm selects relevant features and reduces dimensionality. Initially, data is obtained from the Pima Indian Diabetes Database (PIDD), wherein pre-processing involves handling missing values and normalizing the data by scaling it between the range of 0 and 1. Feature selection using the GWDTO algorithm balances exploration and exploitation to better explore the feature space. The Convolutional Auto encoder (Conv-AE) is then used for classification, after which encoding and decoding of the input data is carried out to capture the underlying structure of the data. The GWDTO algorithm ultimately achieves high accuracy in diabetes diagnosis when compared to the existing techniques namely, Convolutional Neural Networks (CNN) and Naïve Bayes. The proposed method achieves a high accuracy of 99.10% on the PIDD dataset.**

***Keywords—auto encoder, convolutional neural network, deep learning, dipper throated optimization, grey wolf optimization.***

## I. Introduction

Diabetes, a metabolic disorder, is marked by elevated blood sugar levels that result in significant damage to nerves, heart, eyes, kidneys, and blood vessels [1]. This condition gives rise to various associated diseases, marking diabetes among one of the most prevalent health concerns [2]. Analysing a person's information, particularly focusing on two main aspects, can help determine whether they are positive or negative for various diseases, including diabetes. It is crucial for avoiding the missing values and irrelevant columns during the pre-processing phase [3]. For individuals with diabetes, reducing visual impairment is achieved by using feature selection and optimization techniques to enhance performance. These techniques help reduce dimensionality and select relevant features [4-5]. Once the relevant features are extracted, where classification using Deep Learning (DL) techniques can identify health conditions as positive or negative for diabetes based on retinal image screening [6-7].

The main contributions of the paper are discussed below:

- Pre-processing involves the Min-Max technique to scale the data between 0 and 1, along with handling missing values for efficient feature selection.
- The proposed Grey Wolf and Dipper Throated Optimization (GWDTO) algorithm is utilized for feature selection, aiding in the selection of relevant features, balancing exploration and exploitation within the data.
- The Convolutional Auto encoder (Conv-AE) classification is employed to ignore noise, utilizing relevant features efficiently, while achieving high accuracy in classifying diabetes.

The paper is organized as follows: Section 2 provides the related work that summarizes Diabetes by using hybrid optimization techniques, Section 3 introduces the proposed method utilized by GWDTO algorithm, while Section 4 discusses the result and comparative analysis, and Section 5 presents the conclusion of this research.

## II. Literature

Wee et al. [8] presented a DL-based technique using Convolutional Neural Networks (CNN) for diabetes classification using Pima Indian Diabetes Dataset (PIDD). This technique learns from raw pixel data and automatically discovers and adapts to the most salient features of the images such as edges and shapes, classifying efficiently. However, the data's noise leads to ineffective training due to the limited number of data samples and oversampling issues.

Saxena et al. [9] developed a Machine Learning (ML) technique using Random Forest (RF) classifiers for early-stage diabetes classification. This approach reduced the risk of overfitting, handled noise in data, managed missing values, and performed well in high-dimensional spaces when compared to decision trees, resulting in high classification accuracy. However, the complexity and less interpretability of Random Forests when compared to the individual decision trees made the model difficult to understand with a large number of trees causing computational issues.

Olisah et al. [10] introduced an ML framework for diabetes prediction and diagnosis using PIDD, incorporating Twice-Growth Deep Neural Network (2-GDNN) technique. This framework handled large and complex data with

979-8-3503-6066-0/24/$31.00 ©2024 IEEE

connections between layers beginning from the input layer to the hidden layer and then the output layer. It updated the weights using backpropagation to minimize the network error. Nonetheless, it required a large amount of labelled data and overfitting that occurred when the model learnt noise in the training data, rather than the underlying patterns, thereby reducing the accuracy.

Chang et al. [11] implemented a Naïve Bayes to classify data into predetermined categories based on the probability of events, given the dataset as input. This method was computationally efficient in both training and classification, handling large and high-dimensional data quickly. However, Naïve Bayes was significantly affected by data characteristics, requiring categorical input data. Continuous data were needed to be discretized, leading to potential information loss.

Patro et al. [12] implemented a Deep Convolutional Neural Network (DCNN) to enhance the effectiveness of classification methods for accurate diabetes prediction. This technique performed automatic feature extraction, effectively handling large and complex data. Nonetheless, the CNNs required large amounts of labelled training data, which made it difficult and expensive to obtain. Regularization techniques such as dropout layers were needed to handle varying the model sizes.

## III. Proposed Methodology

In this section, the proposed GWDTO algorithm is performed in feature selection to identify relevant features and balance exploration and exploitation in diabetes prediction. The Pima Indian Diabetes (PIDD) dataset is used, and pre-processing involves normalization techniques using Min-Max scaling to scale the data between 0 and 1 and handle missing values. Feature selection is performed using the GWDTO algorithm, and Conv-AE is used for classification to efficiently learn complex features from the input data. Figure 1 shows a block diagram of the proposed method.

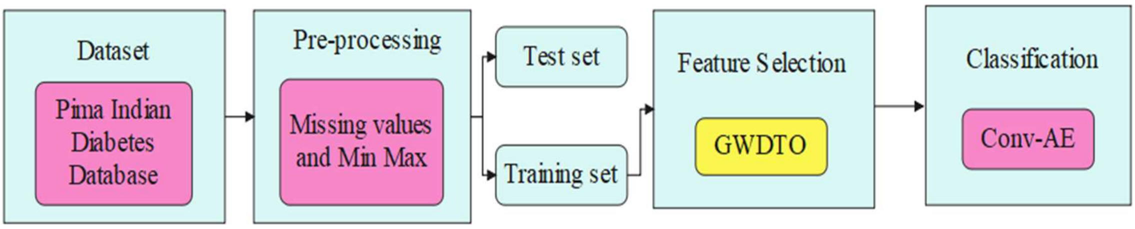

Fig. 1. Block diagram of proposed method

### A. Data Collection

The PIDD [13] is a publicly available dataset used for diabetes prediction and diagnosis for both women and men due to the high number of fatalities associated with diabetes (2.3 million for women and 1.9 million for men). Accordingly, the dataset is used to evaluate the risk of diabetes among affected people based on gender, observing and determining the features selected for diabetes assessment.

### B. Pre-processing

Data pre-processing is an essential step for the PIDD dataset involving cleaning, transforming, and preparing the data to decrease noise and handle missing values. Missing data values are considered using the mathematical equation (1), which replaces missing data with the mean values.

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

In this equation, the pre-processing step involves identifying and handling missing values which are represented mathematically. Normalization ensures that every feature is on the same scale. Min-max scaling is a general model for scaling data between 0 and 1, as shown in equation (2) below:

$$MinMaxScaling = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

Here, $x$ denotes the original values, and $x_{max} - x_{min}$ represent the maximum and minimum values of $x$, respectively.

### C. Feature Selection

After pre-processing data, feature selection is performed using a hybrid optimization technique that combines Dipper Throated Optimization (DTO) and Grey Wolf Optimization (GWO), referred to as DTGWO. This method selects relevant features with the primary goal of maximizing classification accuracy while minimizing the error rate and the features selected. The hybrid optimization balances exploration and exploitation effectively. DTO enhances search ability based on the unique characteristics of birds. It updates the position and velocity of the food search dynamically. Birds in the DTO model are unique among passerines due to their ability to hunt and swim rapidly in a straight line without pauses. They have a short duration to find their position, and prey detection involves diving headfirst into water, even in fast-flowing streams and turning up stones from the bottom, as mathematically expressed in equation (3) to (5).

$$BL_{nd}(t+1) = BL_{best}(t) - C_{1.}|C_{2.}BL_{best}(t) - BL_{nd}(t)| \quad (3)$$

$$BS(t+1) = C_3 BS(t) + C_4 r_1\big(BL_{best}(t) - BL_{nd}(t)\big) + C_5 r_1\big(BL_{Gbest} - BL_{nd}(t)\big) \quad (4)$$

$$BL_{nd}(t+1) = BL_{nd}(t) + BS(t+1) \quad (5)$$

In optimization algorithms, $BL_{nd}(t)$ and $BL_{best}(t)$ represent the recent and best locations of the bird at iteration t. The coefficients $C_{1.}$ and $C_{2.}$ are adaptive values that change during optimization process based on iteration and other solution. The bird's position improve, with $BS(t+1)$ representing the velocity, $r_1$ denotes the arbitrary values in range [0,1], $BL_{Gbest}$ is the global best location, and $C_3$ signifies weight of values, while $C_4$ and $C_5$ are constants.

GWO is a swarm intelligence-based meta-heuristic algorithm that mimics leadership hierarchy and hunting process of grey wolves in nature. The population is essentially divided into 4 groups: alpha, beta, delta, and omega. The goal of GWO is to dominate 2 lower packs, delta and omega, while alpha and beta lead to social intelligence as the main inspiration for GWO algorithm. The hunting approach of grey wolves involves exploration and exploitation, balancing the search ability. It focuses on the best solution found by the alpha wolf, refining it through the social hierarchy, repercussive to the local search around the promising areas. The encircling behaviour is modelled using equations (6) and (7).

$$\vec{F}(t+1) = \vec{F}_p(t) - \vec{A}.\vec{D} \tag{6}$$

$$\vec{D} = \left|\vec{C}.\vec{F}_p(t) - \vec{F}(t)\right| \tag{7}$$

Where, $\vec{A}$ and $\vec{C}$ are vectors of coefficient, and the current iteration is denoted by t, $\vec{F}$ denotes grey wolf location vector and $\vec{F}_p$ denotes the position's vector of prey. The better solution of the iteration and the updated best solution is $\vec{F}$, as formulated in equation (8) to (10).

$$\vec{a} = 2 - t\left(\frac{2}{Max_{iter}}\right) \tag{8}$$

$$\vec{A} = e\vec{a}.\vec{r}_1 - \vec{a} \tag{9}$$

$$\vec{C} = 2\vec{r}_2 \tag{10}$$

Where, the loop count is denoted as t and $Max_{iter}$ indicates maximum number of the iterations of $\vec{r}_1$ and $\vec{r}_2$ arbitrarily vectors within the range of [0,1], and $\vec{a}$ linearly decreases from 2 to 0 iteration length. The exploration is location prey in the 2D search region in the optimum. This consequently maintains 4 search outcomes, as shown in equation (11) to (13).

The GWO is performed forward in 2D search space in 4 phase, locating the grey wolf search space and balancing the exploration and exploitation in the feature. The GWO updates the position in DTO and then efficiently performs feature selection.

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{D_1} * \overrightarrow{F_\alpha} - \vec{F}\right|, \overrightarrow{D_\beta} = \left|\overrightarrow{D_2} * \overrightarrow{F_\beta} - \vec{F}\right|, \overrightarrow{D_\alpha} = \left|\overrightarrow{D_2} * \overrightarrow{F_\delta} - \vec{F}\right| \tag{11}$$

$$\vec{F}_1 = \vec{F}_\alpha - \vec{A}_1 * \vec{D}_\alpha, \vec{F}_2 = \vec{F}_\beta - \vec{A}_2 * \vec{D}_\beta, \vec{F}_3 = \vec{F}_\delta - \vec{A}_3 * \vec{D}_\delta \tag{12}$$

$$\vec{F}(t+1) = \frac{\vec{F}_1 + \vec{F}_2 + \vec{F}_3}{3} \tag{13}$$

### *D. Classification*

The Convolutional Autoencoder (Conv-AE) enhances data classification outcomes by improving classification accuracy through linearly transformed features. Conv-AEs implicitly perform feature selection by learning which features are essential for reconstructing input data. These selected features effectively contribute to learning and achieving high accuracy in classification. AEs learn to ignore noise in input data, thereby improving robustness in classification accuracy. The convolutional layer connects local patterns between adjacent neurons in the layer, storing weights in matrices known as kernels. The convolution operation is defined by equation (14).

$$b_{ij} = f\left((w^k.x)_{ij} + b^k\right) \tag{14}$$

Where, $b_{ij}$ are elements of feature map, $w^k$denotes kernel, x denotes input vector, and $b^k$is bias. The nonlinear function called the activation function is denoted as $f(.)$, operates respectively. The pooling layer performs nonlinear subsampling functions of max-pooling, where the non-overlapping regions are divided to produce maximum outputs. The trained Conv-AE model learns to extract significant attributes of operational status and regenerate values for normal system status. It considers both diabetic and non-diabetic cases where the events are rated by their probable impact on updating weights. During test phase, when encountering situations similar to diabetes, the model identifies them by large reconstruction errors, effectively eliminating disturbances and distortions. In the training phase for diabetes, $D = \{x^1, x^2, \dots, x^d\}$ x represents each training sample evaluated with all variables. The reconstruction error is computed using the equation (15).

$$J = \min_{W,b} \frac{1}{d} \sum_{i=1}^{d} \left|\left|\tilde{x}^{(i)} - x^i\right|\right|_2^2 + \lambda\left|\left|W\right|\right|_2^2 \tag{15}$$

Where, the weights $(W)$ are subject to regularization, characterized by the parameter $\lambda$, $(\lambda \geq 0)$. The cost function in training procedure is the reconstruction error, and stopping condition is achieved by reducing error below a specified threshold after a certain number of iterations. The Convolutional Autoencoder (Conv-AE) operates by mapping data points from beginning to the end through intermediate layers, identifying outliers in the model. Consequently, the errors also exist in diabetic samples due to disturbances caused by varying operating conditions. A threshold is defined to evaluate input samples based on computed reconstruction errors in diabetes. This threshold evaluation aids in assessing reconstruction errors of data during the training process, as shown in equation (16).

$$Thershold = \beta \times mean(Val_{Err}) \tag{16}$$

Where, $\beta$ identifies the level of the diabetes condition and the values of selected features based on a compromise. The Conv-AE reduces dimensionality of data by learning an efficient encoding, capturing the relevant features to aid classification and denoising the data, thereby improving the performance of the classifier.

## IV. Experimental Result

In this research, the proposed LRA technique for classification of diabetes and non-diabetes in PIDD dataset is simulated by using a Python environment with a 16GB RAM, an Intel Core i7 processor, and Windows 10 as operating system. To estimate model's performance metrics namely, accuracy, precision, recall, f1-score and specificity are used. These are numerically formulated in equations (17) and (21).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{17}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{18}$$

$$F1-score = 2 * \frac{(Precision*Recall)}{Precision+Recall} \tag{19}$$

$$Recall = \frac{TP}{TP+Fn} \tag{20}$$

$$Specificity = \frac{TP}{(TN+FP)} \tag{21}$$

Where, $TP, TN, FP$ and $FN$ denote True Positive, True Negative, False Positive, False Negatives, respectively.

### A. Performance Analysis

In this section, the proposed DTGWO-ConvAE method involving feature selection and classification processes is calculated using several performance metrics including Accuracy, Precision, F1-measure, Specificity and Recall for diabetes. The performance of feature selection process with the dataset is represented in Tables 1, which describes the feature selection results. The performance of different classifications with default features using PIDD dataset is represented in Tables 2, which describe the classification results. The performance of GWDTO feature selection is evaluated based on accuracy, precision, F1-measure, Specificity and recall on the PIDD dataset, as described in Table 1. GWO, Whale Optimization Algorithm (WOA), DTO and Particle Swarm Optimization (PSO) are also evaluated. The feature selection technique GWDTO achieves a high accuracy of 99.10%, 97.32% of precision, 97.31% of recall, 97.42% of f1-score and 97.34% of specificity as it selects the related features to easily detect diabetes.

TABLE I. PERFORMANCE ANALYSIS OF THE FEATURE SELECTION IN CONV-AE

| Methods | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Specificity (%) |
|---|---|---|---|---|---|
| PWO | 95.47 | 93.06 | 93.79 | 93.75 | 96.75 |
| GWO | 96.35 | 94.86 | 94.46 | 94.68 | 95.14 |
| WOA | 97.45 | 95.14 | 95.13 | 95.14 | 96.48 |
| DTO | 98.45 | 96.21 | 96.14 | 96.32 | 97.58 |
| GWDTO | 99.10 | 97.32 | 97.31 | 97.42 | 97.34 |

The performance of Conv-AE classification is evaluated based on accuracy, precision, F1-measure, and recall on PIDD dataset, as described in Table 2. The existing methods using classification techniques such as RNN, CNN, DNN, and LSTM are also evaluated. The Conv-AE method achieves a high accuracy of 99.10%, 97.32% of precision, 97.31% of recall, 97.42% of f1-score and 97.34% of on the PIDD dataset. Conv-AE is designed to ignore noise in the input data, leading to more robust classification outcomes when compared to other methods.

TABLE II. PERFORMANCE ANALYSIS OF THE CLASSIFICATION ON PIDD DATASET

| Methods | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Specificity (%) |
|---|---|---|---|---|---|
| RNN | 95.47 | 93.06 | 93.79 | 93.75 | 96.75 |
| CNN | 96.35 | 94.86 | 94.46 | 94.68 | 95.14 |
| DNN | 97.45 | 95.14 | 95.13 | 95.14 | 96.48 |
| LSTM | 98.45 | 96.21 | 96.14 | 96.32 | 97.58 |
| Conv-AE | 99.10 | 97.32 | 97.31 | 97.42 | 97.34 |

### B. Comparative Analysis

The performance of the proposed method GWDTO is compared to existing methods including CNN [8], RF [9], 2GDNN [10], Naïve Bayes [11] and CNN [12]. The comparative analysis involves PIDD datasets where the proposed GWDTO method achieves a high accuracy of 99.10%, 97.32% of precision, 97.31% of recall, 97.42% of f1-score and 97.34% of on PIDD dataset. Table 3 presents a comparative analysis of the proposed method.

TABLE III. COMPARATIVE ANALYSIS OF PROPOSED METHOD

| Methods | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | Specificity (%) |
|---|---|---|---|---|---|
| CNN [8] | 84.37 | NA | NA | NA | NA |
| RF [9] | 79.83 | NA | NA | NA | NA |
| 2GDNN [10] | 99.01 | 97.24 | 97.25 | 97.35 | 97.25 |
| Naïve bayes [11] | 77.83 | 81.23 | 86.09 | 83.60 | 62.03 |
| CNN [12] | 88.38 | 83.33 | NA | NA | NA |
| GWDTO – Conv-AE | 99.10 | 97.32 | 97.31 | 97.42 | 97.34 |

### C. Discussion

This section discusses the advantages of the proposed method and the limitations of the existing methods. The existing method CNN [8] had data noise which caused ineffective training due to the limited number of data samples and oversampling issues. RF [9] possessed complexity and less interpretability of RF when compared to individual decision trees, hence making the model difficult to understand. Additionally, the large number of trees causes computational issues. 2GDNN [10] required a large amount of labelled data and overfitting occurred when the model learned noise in the training data rather than the underlying patterns, thereby reducing the accuracy. NB [11] significantly was affected by data characteristics, requiring categorical input data. Continuous data needs to be discretized, leading to potential information loss. CNNs [12] required large amounts of labelled training data, which was difficult and expensive to obtain. Regularization techniques such as dropout layers needed more capability to handle varying model sizes. The proposed GWDTO algorithm performs feature selection, selecting relevant features and reducing dimensionality. This method enhances accuracy and efficiency in diabetes diagnosis when compared to the existing techniques.

## V. CONCLUSION

This paper proposes a GWDTO algorithm for feature selection to address these challenges. The GWDTO algorithm selects relevant features and reduces dimensionality. Initially, data is obtained from the PIDD, and pre-processing involves handling missing values and normalizing the data by scaling it between 0 and 1. Feature selection using the GWDTO algorithm balances exploration and exploitation to better explore the feature space. The Conv-AE is then used for classification, while encoding and decoding the input data captured the underlying structure of the data. The GWDTO algorithm ultimately accomplishes a commendable accuracy in diabetes diagnosis, as opposed to the existing techniques such as CNN and Naïve Bayes. The proposed method achieves a superior accuracy of 99.10%, 97.32% of precision, 97.31% of recall, 97.42% of f1-score and 97.34% on the PIDD dataset. Future work will explore various classes to improve the method and enhance the accuracy of diabetes prediction.

## References

[1] H. Zhou, Y. Xin, and S. Li, “A diabetes prediction model based on Boruta feature selection and ensemble learning,” BMC Bioinf. , Vol. 24, p.224, June 2023.

[2] S. Gündoğdu, “Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique,” Multimedia Tools Appl, Vol. 82, pp. 34163-34181, March 2023.

[3] S. Gill, and P. Pathwar, “Prediction of diabetes using various feature selection and machine learning paradigms,” In Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough, Part of the book series: Studies in Computational Intelligence ((SCI)), Cham: Springer International Publishing, vol. 1027, pp. 133-146, April 2022.

[4] F. Zia, I. Irum, N.N. Qadri, Y. Nam, K. Khurshid, M. Ali, I. Ashraf, and M. A. Khan, “A multilevel deep feature selection framework for diabetic retinopathy image classification,” Comput. Mater. Contin, Vol. 70, pp. 2261-2276, January 2022.

[5] T. Vijayan, M. Sangeetha, A. Kumaravel, and B. Karthik, “Feature selection for simple color histogram filter based on retinal fundus images for diabetic retinopathy recognition,” IETE J. Res, Vol. 69, pp.987-994, November 2020.

[6] A. A. Alhussan, A. A. Abdelhamid, S. K. Towfek, A. Ibrahim, M. M. Eid, D. S. Khafaga, M. S. Saraya, “Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization,” Diagnostics 2023, Vol. 13, p. 2038, June 2023.

[7] F. Navazi, Y. Yuan, and N. Archer, “An examination of the hybrid meta-heuristic machine learning algorithms for early diagnosis of type II diabetes using big data feature selection,” Healthcare Anal,Vol. 4, p. 100227, December 2023.

[8] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools Appl, Vol. 83, pp. 24153-24185, August 2023.

[9] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, “A novel approach for feature selection and classification of diabetes mellitus: machine learning methods,” Comput. Intell. Neurosci, Vol. 2022, p. 3820360, April 2022.

[10] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” Comput. Methods Programs Biomed, Vol. 220, p.106773, June 2022.

[11] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” Neural Comput. Appl, Vol. 35, pp. 16157-16173, March 2022.

[12] K. K. Patro, J. P. Allam, U. Sanapala, C. K. Marpu, N. A. Samee, M. Alabdulhafith, and P. Plawiak, “An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques. BMC Bioinf, Vol. 24, p. 372, october 2023.

[13] Pima Indian Diabetes Dataset: https://www.kaggle.com/datasets/nancyalaswad90/review (Accessed on July 2024).