iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62867

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Improvement in Prediction to Predict Diabetes for Improved Healthcare

Davinder Singh
UCIM, PU, Chandigarh

Abstract: ***The early prediction and accurate diagnosis of diabetes are critical for effective disease management and treatment. This study presents a novel diabetes prediction model based on a Grasshopper Optimization Algorithm (GOA)-tuned Adaptive Neuro-Fuzzy Inference System (ANFIS). The proposed model aims to enhance accuracy in diabetes prediction by optimizing key parameters within the ANFIS framework using the GOA algorithm. Also, various other issues like complexity, dimensionality, overfitting and under-fitting are mitigated by implementing a hybrid Grey Wold optimization (GWO) and Fuzzy (F) based GWOF Feature selection technique. The efficacy of proposed approach is examined on one of the widely used diabetic datasets i.e., PIMA dataset. Simulations were carried out in MATLAB software under different metrics including accuracy (A), Precision (P), recall (R) and F1-measure (F1-M) on PIMA dataset. The results reveal that proposed GOA-tuned ANFIS model attained an accuracy of 96.22% on given dataset which is higher than standard Non-Linear, DWT, EMD, CNN-LSTM, and CNN-LSTM with SVM considerably.***

Keywords: Diabetes, Glucose Levels, Optimization, Machine Learning etc.

I. INTRODUCTION

Diabetes, a persistent medical condition, is recognized by raised glucose levels beyond safe thresholds because the body cannot properly metabolize the sugar obtained from food. Dysfunction in insulin production and the body's responsiveness to it are both impaired in diabetes [1]. Its symptoms include frequent urination, increased thirst, fatigue, unexplained weight loss, blurred vision, mood swings, difficulty concentrating, and an enhanced susceptibility to frequent infections [2]. Experts estimate that by 2045, there will be 628.6 million diabetics worldwide, up from 424.0 million in 2017. Between 1990 and 2017, the number of fatalities globally attributable to diabetes grew by 0.6 million to 1.7 million [3]. There are three primary forms of diabetes. Type 1, termed insulin-dependent diabetes mellitus (IDDM), is caused by the body's insufficient insulin production, necessitating the use of insulin injections. Type 2, or non-insulin dependent diabetes mellitus (NIDDM), results from the body's cells being unable to utilize insulin effectively. Gestational diabetes, frequently identified as type 3, leads to heightened levels of blood sugar in expecting mothers [4].

Diabetes is considered to arise from a combination of hereditary and environmental influences. There are various risk elements linked to diabetes, which comprise ethnic origin, familial history of diabetes, advancing age, excessive body weight, poor dietary choices, sedentary lifestyle, and tobacco usage [5]. The management of diabetes involves a balance of medication, lifestyle changes, and monitoring. Personalized insulin dosing, regular physical activity, and a balanced diet are vital for blood sugar control. Moreover, advancements in technology, like continuous glucose monitoring (CGM), track glucose levels in real-time, providing continuous data to users, while insulin pumps deliver precise doses of insulin as needed, reducing the risk of blood sugar fluctuations. Additionally, finger stick measurements offer immediate blood glucose readings, facilitating rapid adjustments to treatment plans for optimal control [6]. Despite these measures, long-term poor management of blood glucose levels increases people's vulnerability to illnesses such as heart disease, strokes, kidney failure, neurological impairments, visual difficulties, and more. Therefore, the detection of diabetes in its early stages is imperative, which can allow for proactive medical management, reducing the risk of poor effects through adequate treatment [7].

Diabetes mellitus (DM) can be identified through manual inspection methods, including physical examination by a healthcare provider or analysis of blood glucose levels using traditional laboratory tests. A physical examination for diabetes typically involves assessing vital signs, such as blood pressure, along with examining for symptoms like increased thirst, frequent urination, and unexplained weight loss. Laboratory tests commonly include measuring fasting blood glucose levels, conducting oral glucose tolerance tests, and assessing glycated haemoglobin (HbA1c) levels to evaluate diabetes indicators accurately. However, these manual approaches may exhibit delays in diagnosis, require additional time for interpretation, and are susceptible to human error, which can introduce variability in results, further leading to delays in initiating appropriate treatment [8].

©IJRASET: All Rights are Reserved

Additionally, relying solely on one parameter, whether it's blood pressure, glucose levels, insulin resistance, BMI, or any other factor, for diagnosing diabetes may be ineffective and could result in misleading conclusions during the decision-making process [9]. Artificial intelligence techniques play a crucial role in automating the detection of diabetes, addressing the limitations of manual inspection methods. By effectively mining and analysing data sets, that is, extracting valuable insights from large amounts of information to learn patterns and trends, these techniques lead to improved results [10].

II. LITERATURE REVIEW

We sought to gain an in-depth understanding of the various approaches and machine learning algorithms employed by academics worldwide through a detailed review of the literature that has already been published in the field of diabetes prediction. Khanam, Jobeda Jamal, et al. [11], emphasized early diabetes detection by employing data mining, machine learning, and NN methods on the Pima Indian Diabetes dataset. Among the seven ML algorithms employed and tested, Logistic Regression and Support Vector Machine performed effectively. By implementing the neural network model with varying hidden layers and epochs, they found that the configuration with two hidden layers yielded an accuracy of 88.6%. Roshi Saxena, et al. [12], examined the performance of four ML classifiers (MLP, DT, KNN, and RF) by applying some FS techniques to them with the aim of detecting diabetes in patients at the earliest stages. Moreover, to refine the data, they implemented pre-processing techniques on the original PIMA dataset. The simulations were conducted in Weka 3.9 software, which showed RF got the highest accuracy of 79.8%, while it was only 77%, 76%, and 78% in MLP, DT, and KNN, respectively. Tasin, Isfafuzzaman, et al. [13], developed an automated diabetes prediction system using a proprietary dataset of female patients in Bangladesh alongside the Pima Indian diabetes dataset. The approach utilized mutual information for feature selection and applied extreme gradient boosting with a semi-supervised model to predict insulin features. Moreover, they addressed the class imbalance issue via SMOTE and ADASYN techniques. Additionally, for the classification purpose, they examined the performance of various ML classifiers on given datasets. Results revealed that the XGBoost classifier with the ADASYN approach achieved the highest accuracy of 81%, an F1 coefficient of 0.81, and an AUC of 0.84. Aamir, Khalid Mahmood, et al. [14], proposed an interpretable fuzzy logic model for early diabetes identification, combining it with the cosine amplitude approach to create two fuzzy classifiers and develop fuzzy rules. The model, evaluated on a publicly available diabetic dataset, achieved an accuracy of 96.47%. Abnoosian, Karlo, et al. [15], devised a pipeline-based framework for multi-classification of diabetes across three classes: diabetic, non-diabetic, and prediabetic, by utilizing the imbalanced Iraqi Patient Dataset. The methodology included data pre-processing and FS. Also, multiple ML models (k-NN, SVM, DT, RF, AdaBoost, and GNB) and a weighted ensemble approach were used to address data imbalance issues. To further optimize the performance, they used grid search and Bayesian optimization techniques for hyper-parameter tuning, resulting in superior results with precision (0.9887), recall (0.9861), F1-score (0.9792), AUC (0.9851), and accuracy (0.999).

Hang, Ong Yee, et al. [16], implemented feature selection on a secondary dataset with seventeen attributes and fed it into an AdaBoost with Decision Tree, SVM, and an ensemble model. They identified dataset's top five influential features for training and testing across each model by using SelectKBest function. After comparing predictions from three models, AdaBoost and SVM outcomes were merged to form the ensemble model. The research findings indicated that the ensemble model performs better than individual methods, indicating its suitability for diabetes prediction tools. Ganie, Shahid Mohammad, et al. [17], examined five boosting algorithms on Pima diabetes dataset. They conducted exploratory data analysis to determine the dataset's characteristics, which was followed by upsampling, normalization, feature selection, and hyperparameter tuning for predictive analytics. Results showcased that Gradient boosting attained highest accuracy of 92.85% respectively. Al Sadi, Khoula, et al. [18], examined the performance of six ML algorithms: RF, NB, K-NN, SVM, LDA, and ANN, along with a decision tree, to diagnose T2DM in the Omani population. They obtained the clinical data from a prediabetes register and the Al Shifa health system. A total of 11 clinical features were used to predict T2DM, with RF and DT models exceeding other algorithms by achieving 98.38% accuracy for Oman data and surpassing PID by 9.1% when utilizing the same model and features. Ghane, Sunil, et al. [19], analysed the performance of DT, RF, SVM, KNN, LGBM, and Adaboost on the PID dataset. They considered factors like age, skin thickness, glucose, and insulin to predict diabetes. Through simulations, it was observed that LGBM showed the highest efficiency with 89.85% accuracy and an AUC of 0.95 respectively. Sneha, N., et al. [20], used predictive analysis to select significant features that aid in diabetes mellitus early identification and designed a ML-based prediction algorithm. They employed DT, RF, and NB algorithms on the selected attribute subset. Results showcased that DT and RF showed specificities of 98.20% and 98.00%, respectively, while NB achieved the highest accuracy of 82.30%, generalizing FS to be the main reason behind the enhanced categorization accuracy.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The crux of the above literature is that an ample number of diabetic prediction models have been presented over the last few years. However, certain limitations have been identified, leading us to develop a more improved model. Firstly, it was observed that most conventional approaches use ML classifiers to classify dataset instances; however, these classifiers frequently have issues when dealing with enormous amounts of data.

This is because of their confined computing efficiency, which makes it difficult for models to identify complex data patterns and reduces accuracy. Second, not much work has been done on using effective feature selection techniques, which causes dimensionality issues and ultimately increases the complexity of these models. Keeping these limitations in mind, an improved diabetic prediction model is presented in this manuscript that can mitigate them.

III. PROPOSED WORK

This section of the paper presents an effective diabetic prediction model that is based on soft computing techniques and ML algorithms, focusing on mitigating the limitations of existing ML-based diabetic prediction models. The main objective of the proposed model is to improve the classification accuracy rate while reducing complexity and overcoming dimensionality issues. To attain this objective, work has been done on two important phases of diabetes prediction i.e., Feature Selection and Classification. As observed in the literature section, the prediction accuracy rate can be improved considerably by using an effective FS technique. Keeping this thought in mind, a hybrid GWOF based FS technique is proposed in the first phase of the proposed diabetic prediction model. By employing this technique, an optimal set of features is extracted from the PIMA dataset that aids in enhancing the overall accuracy of the model.

The proposed hybrid GWOF is basically a combination of two techniques i.e., grey Wolf Optimization (GWO) Algorithm and Fuzzy System (FS). One of the primary reasons for using Fuzzy in the proposed work is that it can handle uncertainty in variables like glucose levels and blood pressure measurements in the PIMA dataset by allowing for degrees of truth between 0 and 1, rather than insisting on strict true or false values.

This flexibility enables it to interpret and utilize data points that may not have clear-cut boundaries, accommodating individual differences and measurement errors in the dataset. Moreover, to improve the generalization capability of fuzzy systems GWO has also been used in proposed work.

The GWO algorithm's nature-inspired mechanism allows for better exploration and exploitation, reducing the risk of local minima and improving the accuracy of the diabetes prediction model. However, implementing only hybrid FS technique will not improve the diabetic prediction rate.

Therefore, a GOA-tuned ANFIS classifier has been proposed in the second phase of work to enhance diabetic and non-diabetic categorization accuracy. The ANFIS utilizes the combined capabilities of both fuzzy logic and neural networks, making it well-suited for the classification of instances into diabetic and non-diabetic in dynamic medical datasets like PIMA. Additionally, its capacity to learn from data and improve classification accuracy over time, along with its adaptability to evolving datasets, eventually leads to better generalization. Despite the effective performance of ANFIS due to its numerous advantages, its sensitivity to hyper-parameter settings has great impact on overall accuracy of system. To tackle this challenge, a metaheuristic optimization algorithm named as; GOA is incorporated in proposed work. GOA depicts collective behaviour of grasshoppers to efficiently explore solution spaces for finding optimal solutions. Its ability to rapidly converge and avoid local minima makes it a perfect fit for tuning the hyper-parameters like learning rate, membership function, fuzzy rules etc., of ANFIS. By doing so, it can capture the complex relationships present in the PIMA dataset effectively, which ultimately improves the overall performance of the model. The methodology part of this paper provides a detailed description of the step-by-step operation of the proposed GOA-tuned ANFIS model.

A. Methodology

The several phases that the proposed GOA-tuned ANFIS diabetes prediction model undergoes to ultimately classify diabetic and non-diabetic instances have been briefly discussed in this section.

1) *Collection of Data:* The first phase is to collect the necessary data regarding diabetes and for this, PIMA dataset has been used in proposed work. It is obtained from Kaggle.com and comprises of 768 instances with 8 attributes. Additionally, it also includes a binary target variable depicting diabetes status. The sample of data used in the proposed work has been graphically represented in Figure 1.

©IJRASET: All Rights are Reserved

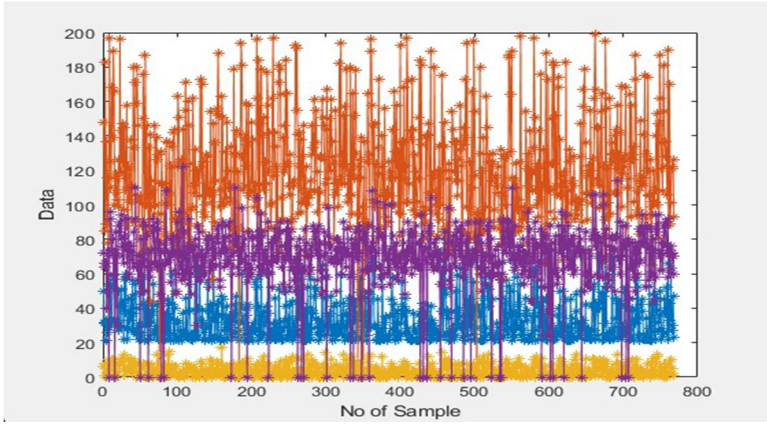


Figure 1. Visualization of the PIMA Dataset

2) *Pre-processing of Data:* Since, the dataset is obtained from internet and may contain lot of null, NAN and unnecessary data that can lead to complexity and decreased accuracy rate. Therefore, pre-processing technique is implemented in proposed work on raw data. During this phase, missing and null values in the PIMA dataset are addressed by substituting them with the average value of the corresponding feature, a technique known as mean imputation. Moreover, the redundant data is also removed from the dataset so that model can get trained effectively on useful data.

3) *Feature Selection:* Once data is processed, it is time to select important features from it to reduce the risk of overfitting, under-fitting, complexity and dimensionality. For this, hybrid GWOF is used in proposed work that effectively selects only those features from the given data which have huge impact on accuracy rate. The initialization parameters of hybrid GWOF are shown in Table 1. Once the hybrid GWOF initialization, three crucial factors, namely standard deviation, mean, and correlation of each feature are obtained which serve as inputs for the fuzzy logic system. These parameters are then processed as per the predefined rules in fuzzy system to derive a singular output weightage, specifying the influence of each feature on the final prediction score. To further optimize the process of FS, GWO has been used in proposed work which analyse the weightage output and selects only those features whose weightage output is high.

Table 1. Initialization Parameters of Hybrid GWOF

Parameter	Value
Max_iter	50
NoSelection	4
LB (Lower Bound)	1
UB (Upper Bound)	(Data,2)
Population Size	10

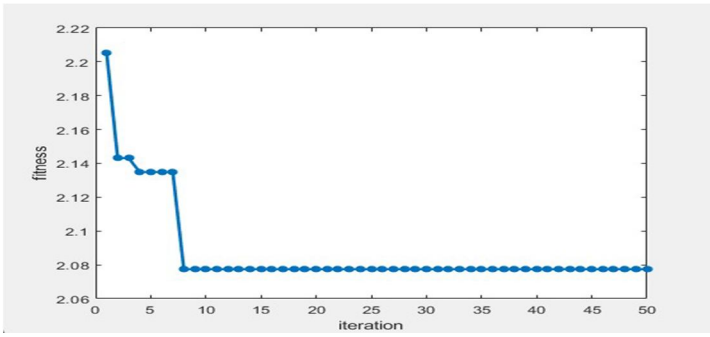


Figure 2. Fitness Convergence Graph for Hybrid GWOF Algorithm

The effectiveness of proposed hybrid GWOF is determined by its convergence curve, depicted in Figure 2. This graph clearly shows that convergence curve of proposed hybrid GWOF gets stabilized after performing just 7^{th} iteration and then remains constant till 50^{th} iteration, showcasing the effectiveness of proposed FS technique.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) *Split Data:* After creating a final list of important features, the data is categorized into two categories of training and testing in the ratio of 70:30 respectively.
5) *Classification:* The training data is then used for training the proposed classifier. Here, we have used ANFIS as base classifier for learning the patterns from given data. However, due to its sensitivity to hyper-parameter setting, the performance can be impacted greatly. Therefore, GOA has been also used in proposed work for tuning the parameters of ANFIS, minimizing the likelihood of encountering local optima. The initialization parameters of GOA are given below.

Table 2. Initialization Parameters of GOA-ANFIS

Parameter	Value
Max Iteration	100
Population Size	30
Lower Bound	-25
Upper Bound	25

The GOA updates the hyper-parameters of the ANFIS, and the model trains on training data. Once the model is trained, testing data is passed to it, to check its efficiency on unseen data. The classifier extracts best features from the data and classifies the instances based on patterns to determine if they are diabetic or not.

6) *Performance Evaluation:* Finally, performance of proposed GOA-tunes ANFIS model is examined for various parameters whose discussion is given in next section.

IV. RESULTS OBTAINED

The performance of the proposed GOA-tuned ANFIS classifier has been assessed by simulating it in MATLAB software to obtain a confusion matrix and the graph of evaluation metrics (Accuracy, F1 score, Precision, and Recall) for it. Moreover, to prove the enhanced classification performance of our proposed model, its accuracy rate is compared with the 5 conventional classifiers (Non-Linear, DWT, EMD, CNN-LSTM, and CNN-LSTM with SVM). All the outcomes that we have acquired are detailed in this section of the paper.

A. Performance Evaluation

To begin the analysis of the performance of the proposed GOA-tuned ANFIS model is examined on PIMA dataset under four different metrics, whose graph is given in Figure 3. From the given graph, it is clear that the accuracy of the proposed GOA-tuned ANFIS model has reached 0.96, accompanied by an F1 score of 0.97. Furthermore, it has been delivering precision and recall values of 1.00 and 0.95, respectively, affirming its proficiency in correctly classifying diabetic and non-diabetic instances.

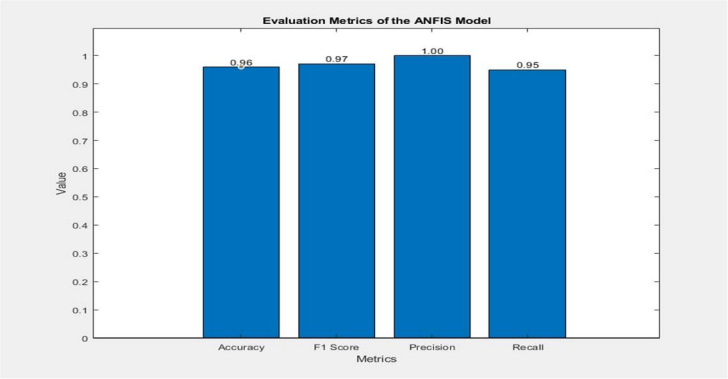


Figure 3. Evaluation Metrics for the proposed ANFIS Model

Furthermore, to prove the efficacy of proposed approach over conventional diabetic prediction models, we have done a comparison analysis centered on an accuracy metric in order to verify the superiority of the proposed-ANFIS-GOA model. The results that we obtain are shown in Figure 4.

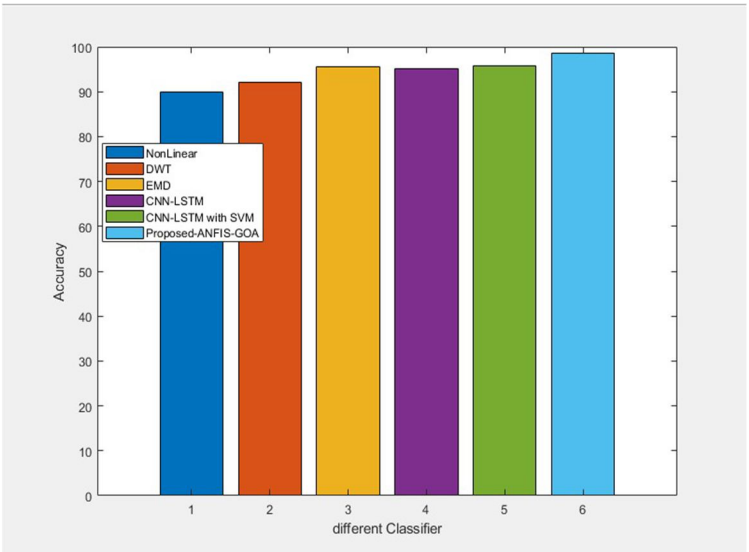


Figure 4. Comparative accuracy of varied classifiers

From the given graph, it is clear that out of all models, standard Non-Linear model has lowest accuracy of 90%, while as it is around 92% in DWT, 95% in EMD, 95.1% in DL based CNN-LSTM and 95.7% in DL based CMM-LSTM with SVM respectively. On the contrary side, the accuracy rate attained in proposed GOA-ANFIS model came out to be 96.22% on PIMA dataset which means there is an improvement of around 0.52% in accuracy of proposed approach when compared with best performing traditional model. Table 3 lists the precise numerical values obtained for each model.

Table 3. Accuracy Values for different classifiers

Classifier	**Accuracy (%)**
Non Linear	90
Discrete Wavelet Transform	92
Empirical Mode Decomposition	95.63
Deep Learning (CNN-LSTM)	95.1
Deep Learning (CNN-LSTM with SVM)	95.7
Proposed GOA-ANFIS	96.22

From the above given graphs and tables, it is clear that proposed GOA-tuned ANFIS mode is outperforming other given traditional diabetic prediction models by attaining an accuracy of 96.22%. The results showcase that largest improvement is observed compared to the Non Linear classifier, with an increase of 6.22%. The smallest improvement is compared to the Deep Learning (CNN-LSTM with SVM), with an increase of 0.54% respectively. These high value of accuracy clearly generalizes and validates the role of Hybrid GWOF feature selection and the GOA-tuned ANFIS model for improved and accurate diabetic prediction.

V. CONCLUSION

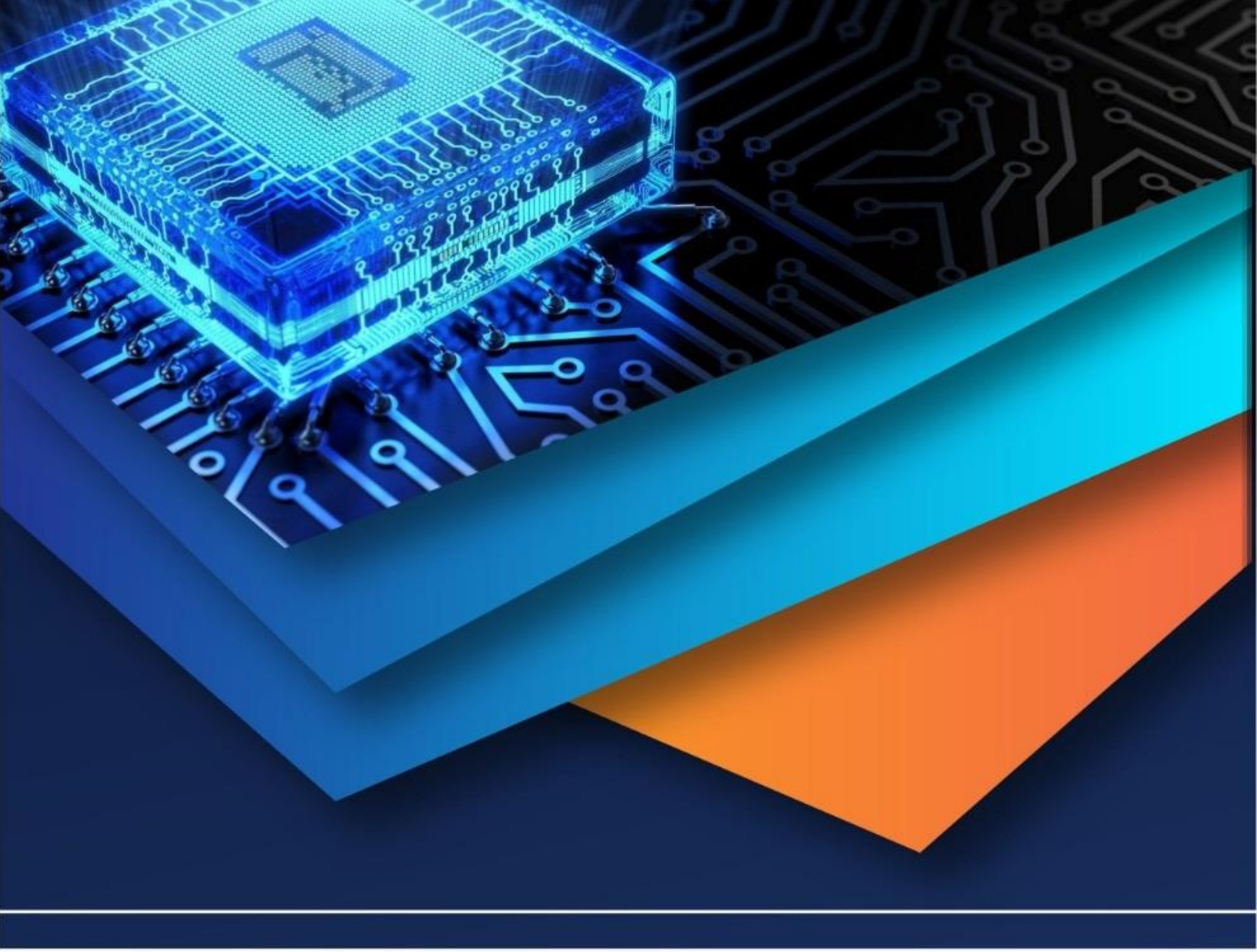
This paper presented an efficient and accurate diabetic prediction model, named as GOA-tuned ANFIS, whose simulations are performed in MATLAB software. The simulating results are observed on PIMA dataset, which is one of the widely used dataset for conducting research in diabetes prediction. Through the results, we observed that proposed GOA-tuned ANFIS model attains highest accuracy of 96.22% on given dataset to prove its effectiveness over other similar models. Upon comparing this accuracy rate with traditional models, we observed that there is an improvement of 6.22%, 4.22, 0.59%, 1.12% and 0.52% in proposed model when compared to Non-Linear, DWT, EMD, DL based CNN-LSTM and DL based CNN-LSTM with SVM respectively. These results suggest that the GOA-tuned ANFIS model provides a more accurate approach to diabetes prediction, making it a promising approach for further development and practical application in medical diagnosis and patient care. The improvements across a diverse set of baseline classifiers demonstrate the robustness and versatility of this proposed model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Naz, Huma, and Sachin Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset". Journal of Diabetes & Metabolic Disorders 19 (2020): 391-403.
[2] Rani, K. J. "Diabetes prediction using machine learning". International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305.
[3] Reza, Md Shamim, et al. "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data". Heliyon 10.2 (2024)
[4] AC, Ramachandra, and Dhanush Murthy. "Diabetes Prediction Using Machine Learning Approach". (2023)
[5] Chang, Victor, et al. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms". Neural Computing and Applications 35.22 (2023): 16157-16173
[6] Sempionatto, Juliane R., Jong-Min Moon, and Joseph Wang. "Touch-based fingertip blood-free reliable glucose monitoring: Personalized data processing for predicting blood glucose concentrations". ACS sensors 6.5 (2021): 1875-1883
[7] Ahmed, Nazin, et al. "Machine learning based diabetes prediction and development of smart web application". International Journal of Cognitive Computing in Engineering 2 (2021): 229-241.
[8] Chaki, Jyotismita, et al. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review". Journal of King Saud University-Computer and Information Sciences34.6 (2022): 3204-3225.
[9] Alam, Talha Mahboob, et al. "A model for early prediction of diabetes". Informatics in Medicine Unlocked 16 (2019): 100204
[10] Gujral, Sakshi. "Early diabetes detection using machine learning: a review". Int. J. Innov. Res. Sci. Technol 3.10 (2017): 57-62.
[11] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction". Ict Express 7.4 (2021): 432-439.
[12] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods". Computational Intelligence and Neuroscience, vol. 2022, Article ID 3820360, 11 pages, 2022.
[13] Tasin, Isfafuzzaman, et al. "Diabetes prediction using machine learning and explainable AI techniques". Healthcare technology letters 10.1-2 (2023): 1-10.
[14] Aamir, Khalid Mahmood, et al. "A fuzzy rule-based system for classification of diabetes". Sensors 21.23 (2021): 8095.
[15] Abnoosian, Karlo, Rahman Farnoosh, and Mohammad Hassan Behzadi. "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models". BMC bioinformatics 24.1 (2023): 337.
[16] Hang, Ong Yee, Wiwied Virgiyanti, and Rosly Rosaida. "Diabetes Prediction using Machine Learning Ensemble Model". Journal of Advanced Research in Applied Sciences and Engineering Technology 37.1 (2024): 82-98.
[17] Ganie, Shahid Mohammad, et al. "An ensemble learning approach for diabetes prediction using boosting techniques". Frontiers in Genetics 14 (2023): 1252159.
[18] Al Sadi, Khoula, and Wamadeva Balachandran. "Prediction model of type 2 diabetes mellitus for Oman prediabetes patients using artificial neural network and six machine learning Classifiers". Applied Sciences13.4 (2023): 2344.
[19] Ghane, Sunil, et al. "Diabetes Prediction using Feature Extraction and Machine Learning Models". 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2021.
[20] Sneha, N., Gangil, T. "Analysis of diabetes mellitus for early prediction using optimal features selection". J Big Data 6, 13 (2019).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)