Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Optimizing Random Forest for Type II Diabetes Prediction Using Firefly Algorithm

Michael Favour Edafeajiroke,
Department of Computer Science
University of Port Harcourt
Rivers State Nigeria

Veronica Ijebusomma Osubor,
Department of Computer Science
University of Benin
Edo State, Nigeria

***Abstract* - This study develops an enhanced Random Forest (RF) model for Type II diabetes prediction, optimized using the Firefly Algorithm (FA) for feature selection. The aim was to address the limitations of standalone RF models, which often suffer from suboptimal performance without intelligent feature selection and parameter tuning. A standardized preprocessing pipeline using StandardScaler was implemented on a clinical dataset. The FA was then employed to identify an optimal feature subset, which was used to train the RF classifier. The hybrid FA-RF model achieved a superior accuracy of 99.91%, along with a sensitivity of 98.96%, specificity of 100%, and precision of 99.95%, significantly outperforming the baseline RF model (97.06% accuracy, 69.15% sensitivity). Results demonstrate that integrating metaheuristic optimization with ensemble learning creates a more accurate and reliable tool for early diabetes prediction. The model's high performance and robustness highlight its strong potential for deployment in real-world clinical settings as a decision-support system to improve screening and patient outcomes.**

Keywords: Type II Diabetes Prediction, Random Forest, Firefly Algorithm, Feature Selection, Machine Learning, Clinical Decision Support.

I. INTRODUCTION

Type II Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from insulin deficiency or resistance. It represents a significant and growing global health challenge; according to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, a figure that is expected to increase to 643 million by 2030 [1]. This increasing prevalence underscores the urgent need for early and accurate prediction to mitigate severe complications such as cardiovascular diseases, kidney failure, and neuropathy [2, 3]. Traditional diagnostic methods, which rely heavily on clinical tests and physician expertise, are often time-consuming, expensive, and prone to human error [4]. Consequently, machine learning (ML) techniques have emerged as powerful tools for enhancing the accuracy and efficiency of diabetes prediction [5].

Among these, the Random Forest (RF) algorithm has gained prominence due to its robustness in handling high-dimensional data and reducing overfitting [6]. However, despite its effectiveness, recent studies indicate that the performance of RF can be suboptimal without careful feature selection and hyperparameter tuning, which can limit its predictive accuracy and clinical utility [7, 5]. Diabetes datasets often contain a mix of clinical and lifestyle variables, where irrelevant or redundant features can obscure critical risk factors and lead to models that are computationally inefficient and less interpretable for healthcare professionals. This creates a specific need for intelligent optimization tailored to the heterogeneous nature of medical data.

To address these limitations, nature-inspired optimization algorithms such as the Firefly Algorithm (FA) have shown considerable promise in improving ML models by optimizing feature subsets [8]. The FA, which mimics the flashing behavior of fireflies to solve complex optimization problems efficiently, has been successfully applied in various domains of medical diagnostics [9]. Its particular strength lies in efficiently navigating high-dimensional search spaces to identify parsimonious, clinically-relevant feature subsets, which can enhance model performance and interpretability a crucial factor for clinical adoption [10]. Furthermore, recent research highlights the growing adoption of such metaheuristic algorithms in healthcare analytics for improved decision-making [11, 12].

Despite these advancements, a notable research gap persists. While metaheuristics are applied in healthcare, there is limited focused research on integrating the FA specifically with RF for Type II diabetes prediction to simultaneously boost accuracy and model interpretability through optimal feature selection [13]. Existing studies often focus on standalone classifiers or complex hybrids without leveraging the FA's efficiency for feature selection within the robust RF framework [14]. Moreover, many current AI models lack rigorous validation of their generalizability across diverse datasets, raising concerns about their clinical applicability [15].

Therefore, this study aims to bridge this gap by proposing a hybrid FA-RF model specifically designed for Type II diabetes prediction. The integration is particularly beneficial as the FA optimizes the feature subset to reduce dimensionality and highlight key diagnostic variables, thereby enhancing the RF model's accuracy, speed, and clinical interpretability. The specific objectives of this research are to:

1. preprocess a clinical diabetes dataset using a standardized StandardScaler technique.
2. perform feature selection on the preprocessed dataset using the Firefly Algorithm.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

3. develop a Random Forest model using the optimal feature subset and evaluate its performance using metrics including Accuracy, Precision, Sensitivity, Specificity, F1-score, and AUC.

4. compare the performance of the FA-optimized RF model against a baseline RF model without FA-based feature selection.

II. LITERATURE SURVEY

A consistent theme across the literature is the pursuit of higher predictive accuracy using diverse algorithms. Studies such as [16] and [17] demonstrated the potential of models like K-Nearest Neighbors (KNN) and CatBoost, achieving accuracies of 96% and 95.4%, respectively. Similarly, advanced hybrid and optimization techniques have shown remarkable performance. For instance, [18] proposed a novel feature selection method achieving 98.6% accuracy, while [19] used a Grey Wolf Optimized MLP to reach 97% accuracy. [20] and [21] specifically highlighted the promise of the FA for feature selection and model optimization, with the latter achieving an Area Under Curve (AUC) of 0.98 for predicting patient readmission. However, these studies collectively reveal several critical limitations that this research seeks to address. A primary issue is the inconsistency in data preprocessing across comparative analyses. Many studies, including those by [16] and [22], compared multiple classifiers but lacked a unified preprocessing pipeline. This omission makes it difficult to ascertain whether performance differences stem from the algorithms' inherent capabilities or from inconsistencies in input scaling and transformation, thereby compromising the fairness and reproducibility of the results. Many studies exhibit a narrow scope of prediction. For example, [23] focused exclusively on cognitive decline in diabetic patients, rather than the core task of diabetes prediction itself. This specialization limits the generalizability of their findings to broader screening contexts. Furthermore, a common weakness is the presence of demographic bias in datasets, as noted by [17], which can lead to models that perform poorly on populations not represented in the training data. While powerful, many sophisticated optimization approaches [18, 19, 24] are highly complex, and their comparative advantage over a rigorously tuned, feature-optimized standard classifier remains unclear. Conversely, studies that employed simpler models often omitted robust feature selection, as seen in [25], which acknowledged that its Random Forest model's efficiency could be improved with optimal feature subset selection. This gap between complex optimizers and simpler models presents an opportunity for a streamlined yet powerful approach.

This study was designed to mitigate these identified gaps. Unlike previous works, it implemented a standardized preprocessing strategy using StandardScaler to ensure consistent input scaling, harnessing the inherent ensemble efficacy of the random forest model, facilitating a fair and reproducible comparison. Tackling dataset bias and enhancing model efficiency, this study further integrated the FA for feature selection, building upon its promising applications shown by [20, 21]. Rather than developing a new complex hybrid, this research focused on enhancing the robust and interpretable Random Forest algorithm by leveraging the FA to identify the most predictive feature subset. Finally, the model was evaluated on core diabetes prediction, ensuring broad applicability, and its performance was rigorously assessed using a comprehensive set of metrics, including Accuracy, Precision, Sensitivity, and Specificity, as well as F1 score, AUC score/ROC Curve. By integrating a standardized preprocessing pipeline with a metaheuristic-optimized feature selection process for a powerful yet interpretable algorithm, this research aims to produce a more generalizable, accurate, and clinically applicable model for diabetes prediction.

III. METHODOLOGIES

A. Research Methods

The experimental framework of the study is expressly stated in sequential steps in view of achieving the stated objectives. By using the fireflies' flashing patterns and behaviors to optimize the type II diabetes features obtained from the Kaggle repository dataset from https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset. This dataset was selected for its clinical relevance, standard use as a benchmark in diabetes prediction research, and comprehensive inclusion of key diagnostic and lifestyle features. While other public datasets like the PIMA Indians Diabetes Database exist, the chosen dataset provides a larger sample size (100,001 instances) and a modern feature set including HbA1c_level, which is a critical contemporary diagnostic marker. This enhances the model's relevance to current clinical practice. Hence, with this dataset, the experiments aim to improve the random forest model's performance. The optimal feature subset obtained was used to build a random classifier model for predicting type II diabetes in patients. The system was designed to get an optimal subset from the dataset so as to avoid outliers, under-sampling, and oversampling in the dataset, designed to simulate a data mining sequential process. Random Forest Classifier has been a good classification and regression algorithm, but to get a more efficient result, the FA was utilized for the study. The FA was used in a wrapper approach to ascertain the pertinent type II diabetes features subset for improved effective performance of the random forest model during the type II diabetes prediction process.

Fig 1. provides a workflow model for the sequential steps involved for actualize the study aim. This includes the collection diabetes dataset from the Kaggle dataset repository. The FA was employed for feature selection by optimizing the high-dimensional dataset for better performance of the random forest model. Also, the preprocessed features were passed to the Random Forest without utilizing the FA, to ascertain the effect of the FA on the performance of the random forest. Utilizing the FA for features helped to identify the main factors and the lifestyle responsible for Type II diabetes, and the best features were then used to develop a Random forest model. The dataset was partitioned into two folds: the training and testing sets. Results obtained from each developed model were compared with each other to ascertain the best model for diabetic prediction on the dataset. The training set and testing set of data were divided at a percentage ratio of 75% to 25%, respectively. The results were evaluated based on ML statistical metrics like the classification accuracy, specificity, sensitivity, and precision. This process was implemented on a Python Jupyter Notebook on a high-performance platform for a Machine Learning Platform.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

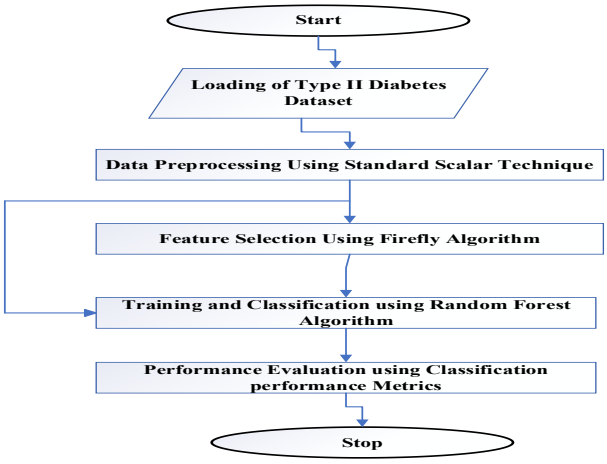

Fig 1. Workflow Diagram for the Developed Type II diabetes prediction model

B. Procedural Techniques for the Developed System

The methodology this study adopted to achieve the study aim was a combined Knowledge Discovery in Databases (KDD) and Machine learning design methodology. The study utilized two key techniques, the FA and the random forest algorithm. The data mining techniques adapted are shown in this order, from data acquisition, through feature selection, classification, and then evaluation of the different models.

i. Dataset acquisition

A dataset was used to formulate a knowledge base for the system so as to build an efficient system. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The Dataset consists of nine (9) different attributes, of which one (1) was taken as the target variable or the class label, and then Ten Thousand and one responses (100001). This dataset was preprocessed with the standard scalar technique before utilizing the FA for the feature selection stage. Attribute description of the various dataset is shown in Table 1.

ii. Data Set Pre-Processing and Normalization

Removing errors and outliers that may be present in the data are components of the pre-processing task that should be performed to make the information appropriate for modeling. In order to avoid inconsistency, imbalance, and missing responses normally prone to diabetes datasets in order to achieve the above, StandardScaler to normalize the dataset. StandardScaler transforms data to have a mean of zero and a standard deviation of one.

iii. Feature Selection using the Firefly Algorithm

To address dimensionality and identify the most predictive feature subset, the Firefly Algorithm (FA) was employed in a wrapper-based approach. The FA is a nature-inspired metaheuristic that optimizes feature subsets by simulating the attractiveness of fireflies based on their brightness (fitness). The fitness function for this study was defined as the classification accuracy of a preliminary Random Forest model using the selected feature subset. The algorithm parameters were set as follows: number of fireflies=20, maximum generations=100, absorption coefficient=1.0, and randomization parameter=0.2. This process aimed to reduce noise, lower computational cost, and enhance model interpretability by identifying the most clinically significant risk factors, with the optimal subset presented in Table 3.

Table 1. Type II Diabetes Feature Description

Feature Index	Feature Name
1	(string) Gender
2	(int) Age
3	(bool) hypertension
4	(bool) heart_disease
5	(string) smoking_history
6	(Float) bmi
7	(Float) HbA1c_level
8	(int) blood_glucose_level
9	(bool) diabetes: target variable or Class Label

iv. Training and Classification

The preprocessed dataset was partitioned into training (75%) and testing (25%) sets. This split provides a substantial amount of data for learning while retaining a robust hold-out set for unbiased evaluation. To further ensure the robustness and generalizability of the results, a 10-fold cross-validation procedure was applied during the model training and tuning phase on the training set. The FA-optimized feature subset was used to train the Random Forest classifier. For comparison, a baseline RF model was also trained using all features post-scaling, but without FA-based selection. The Random Forest algorithm was implemented with 100 estimators (trees) and the Gini impurity criterion.

C. Architectural Framework for the Developed System

The framework for the developed system is described in Fig 2. The model gives a detailed description of how the study's aim and objectives were realized. Cascading processes of the various blocks are depicted. The first stage shows the input block called the dataset block containing 9 attributes. The output from the input block was fed to the preprocessing stage, where the StandardScaler technique was employed, and the 9 attributes were reduced by sending it to the third block, which is the feature selection stage. An optimal feature subset obtained from the FA was sent to the fourth stage, which is the classification stage, where an RF classification model was built and used to classify the diabetic dataset. To achieve this diabetes dataset was partitioned into 75% for training and 25% for testing. Based on the user input, the overall output stage was able to detect if a patient was diabetic or not diabetic. User friendliness and easy operation were achieved. Based on the results obtained, recommendations were made to patients for treatments.

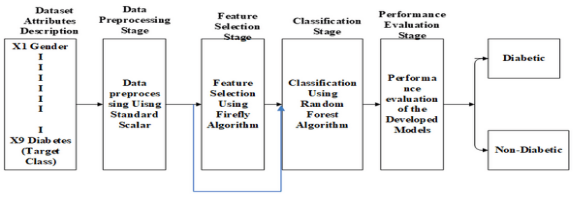

Fig 2. Architectural Framework for the Developed Type II Diabetes System

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

D. Type II Diabetes Performance Evaluation Metrics

Model performance was evaluated using a comprehensive set of metrics derived from the confusion matrix (True Positives-TP, True Negatives-TN, False Positives-FP, False Negatives-FN) [28]. Table 2 presents the equations for the primary metrics:

Table 2. Performance Evaluation Metrics

Measure	Formula
Precision	$\frac{TP}{TP + FP}$ (1)
Sensitivity	$\frac{TP}{TP + FN}$ (2)
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$ (3)
Specificity	$\frac{TN}{TN + FP}$ (4)

Additionally, the F1-score (harmonic mean of precision and recall) and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were calculated. Sensitivity is of paramount importance in a medical screening context to minimize false negatives, where a diabetic patient is incorrectly predicted as healthy. The workflow of the entire methodology is summarized in Figure 1, and the system architecture is detailed in Fig 2.

IV. RESULTS AND DISCUSSIONS

In this section of this paper, the findings derived after implementation were presented. The paper aimed at optimizing the performance of the Random Forest (RF) model by utilizing the efficacy of the flashing behaviors of the firefly to formulate an optimization algorithm that was, in turn, utilized for optimizing the already processed type II diabetes.

In this research, the Python programming language was used to implement all the techniques and models. The environment used for the execution of each program code was the Jupiter notebook, which was built specifically to run Python code, as shown in Fig 3. The models were put into practice on the cloud-based platform Google Colab, which is specifically well-suited for the prediction of type II diabetes. The experimental computer utilized for the simulation is shown in Fig 4. It is a Dell notebook PC, model name Latitude E7470, with an Intel(R) Core(TM) i7-6600U CPU @ 2.60GHz, 2801 Mhz, 2 Core(s), 4 Logical Processor(s), and 16GB of system RAM.

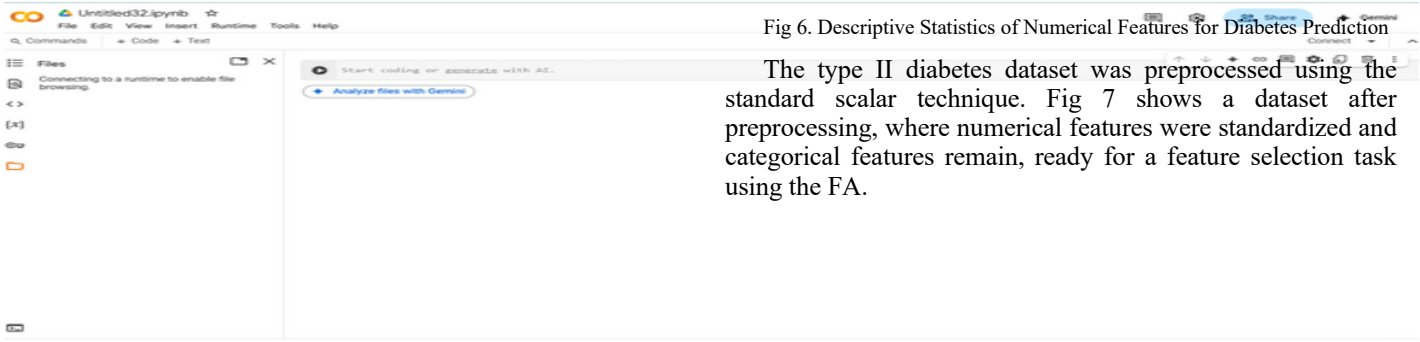

Fig 3. Python Command Window

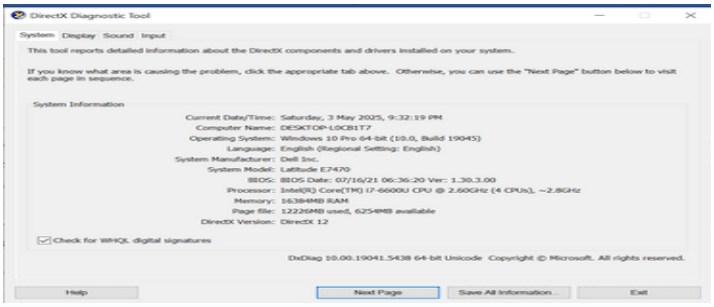

Fig 4. Configuration of the Experimental Machine

Fig 5. provides a Preview of the Patient Health Data for Type II Diabetes Analysis. It describes the pandas DataFrame with 100,000 Type II diabetes patient records and 9 health-related columns or features. The data includes both numerical measurements (like age, BMI, and glucose levels) and categorical information (like gender and smoking history), with no missing values in any column.

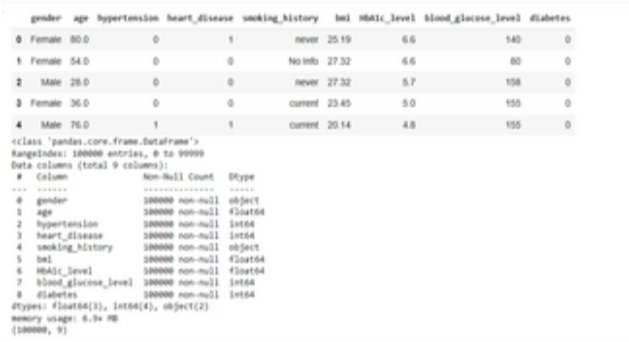

Fig 5. Preview of Patient Health Data for Type II Diabetes Analysis

This statistical summary from Fig 6 reveals key risk factors for diabetes prediction: the average HbA1c level (5.53%) and blood glucose (138 mg/dL) are notably high, hovering near pre-diabetic thresholds. The data show that while the overall diabetes prevalence is 8.5%, the wide ranges in glucose, HbA1c, and BMI indicate a population with significant metabolic diversity, which is crucial for building a predictive model.

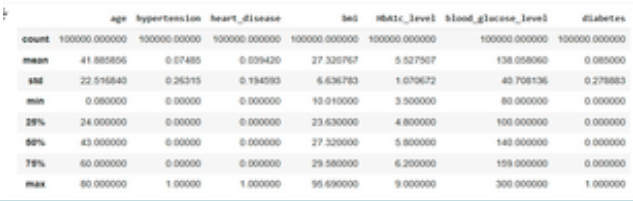

Fig 6. Descriptive Statistics of Numerical Features for Diabetes Prediction

The type II diabetes dataset was preprocessed using the standard scalar technique. Fig 7 shows a dataset after preprocessing, where numerical features were standardized and categorical features remain, ready for a feature selection task using the FA.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	gender	smoking_history	diabetes
0	1.692704	-0.284439	4.936379	-0.321056	1.001706	0.047704	Female	never	0
1	0.538006	-0.284439	-0.202578	-0.000116	1.001706	-1.426210	Female	No Info	0
2	-0.616691	-0.284439	-0.202578	-0.000116	0.161108	0.489878	Male	never	0
3	-0.261399	-0.284439	-0.202578	-0.583232	-0.492690	0.416183	Female	current	0
4	1.515058	3.515687	4.936379	-1.081970	-0.679490	0.416183	Male	current	0

Fig 7. Preprocessed Type II Diabetes Dataset for Feature Selection

The acquired type II diabetes dataset had 9 features after optimization with the FA. Out of these nine (9) features, six (6) features were selected this is as presented in Table 3. Hence FA identified the most critical factors for predicting type II diabetes, which are a combination of lifestyle choices and key clinical biomarkers. The model achieved an excellent accuracy of nearly 96% and a training time of 0.416 seconds, but a longer time was taken to select all features, as presented in Table 4. Hence FA has been demonstrated to be a highly effective and efficient feature selection technique.

Table 3. Optimal Type II diabetes Feature Subset from the Firefly Algorithm

S/N	Selected Features
1	smoking_history_current
2	smoking_history_not current
3	hypertension
4	bmi
5	HbA1c_level
6	blood_glucose_level

Table 4. Model Performance Metrics Comparison as Obtained from the Firefly Algorithm

S/N	Model	Accuracy	Training Time (s)
0	All Features	0.95890	0.504735
1	Selected Features (FA)	0.95955	0.415618

Fig 8. visually shows the specific health factors identified by the FA as most important for predicting Type II diabetes. The length of each bar represents how frequently each feature was chosen as a key predictor. Fig 9 visually presents the training and testing time, and Fig 10 is the conceptual illustration of the FA brightness fitness. The visualizations confirmed that the selected features were predominantly related to smoking history, hypertension, BMI, and key diabetes indicators (HbA1c and blood glucose).

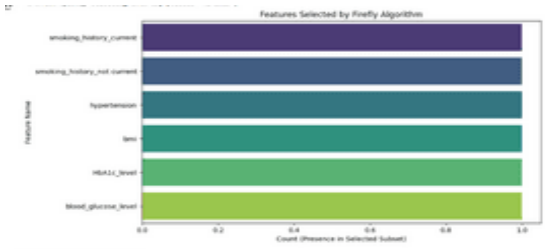

Figure 8. Feature importance as obtained from the Firefly Algorithm

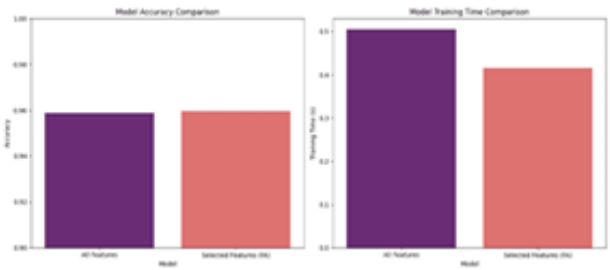

Fig 9. Training and testing Time for the Firefly Feature Selection technique

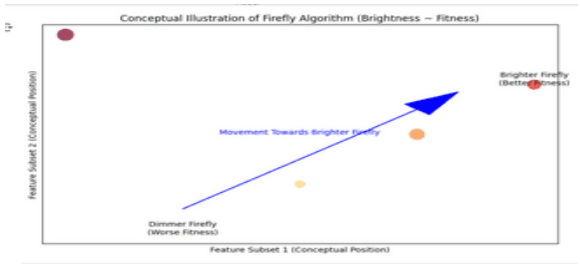

Fig 10. Conceptual Illustration of the Firefly Algorithm Brightness Fitness

Table 5 presents a comparative analysis of two distinct approaches when applied alongside a Random Forest machine learning algorithm. The first technique employs a combination of Standard Scalar and Feature optimization technique with FA before model training, while the second technique utilizes Standard Scalar alone. A clear and significant performance differential is immediately evident between these two approaches. The integrated Standard Scalar and FA method achieves an exceptional accuracy rate of 99.91 percent. Furthermore, this technique demonstrates near-perfect precision and specificity scores of 99.95 and 100 percent, respectively. It also maintains a remarkably high sensitivity rate of 98.96 percent, indicating its proficiency in identifying true positives. Consequently, its F1 and AUC scores are both exceptionally strong at 99.44 and 96.90 percent, confirming a robust and well-balanced model. In stark contrast, the method relying solely on Standard Scalar attains a notably lower accuracy of 97.06 percent. This approach exhibits a critical deficiency in its sensitivity, which plummets to just 69.15 percent, revealing a high rate of false negatives. Therefore, the conclusion is unequivocal that the incorporation of FA as a Feature selection technique is a vital step for optimizing the model's predictive power and overall efficacy.

Table 5. Comparative Analysis of the Developed Type II Diabetes Prediction Model

Technique	Accuracy (%)	Precision (%)	Specificity (%)	Sensitivity (%)	F1 Scores (%)	AUC Scores (%)
Standard Scalar + FA+ RF	99.91	99.95	100	98.96	99.44	96.90
Standard Scalar + RF	97.06	95.17	99.67	69.15	95.45	95.21

A comparative analysis of the two models' accuracy was pictorially illustrated in Fig 11. Results from the model using Feature selection (FA) achieved higher accuracy (99.91%) than the model without it (97.06%).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

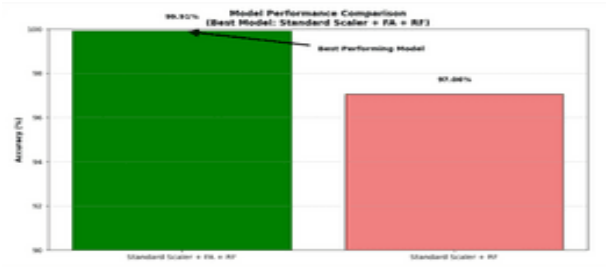

Fig 11. A comparative analysis of the classification accuracy of Two Random Forest Modeling Approaches for Type II Diabetes

An illustration comparing the precision obtained from the two Random Forest Modeling Approaches for Type II Diabetes is presented in Fig 12. The Standard Scalar + FA + RF technique achieved near-perfect precision at 99.95%, significantly outperforming the Standard Scalar + RF method, which scored 95.17%. This demonstrates that adding Feature selection (FA) provides a substantial boost to the model's accuracy in predicting Type II diabetes.

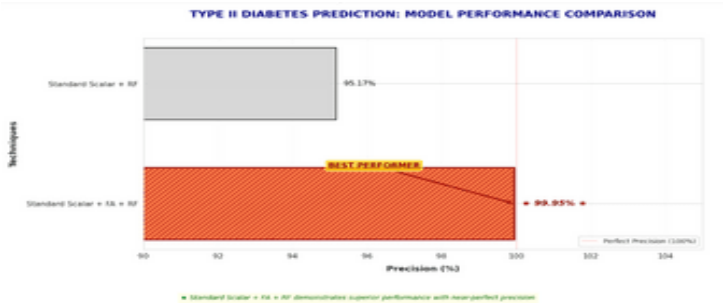

Fig 12. A comparative analysis of the Precision of the two Random Forest Modeling Approaches for Type II Diabetes

Fig 13 provides a pictorial illustration of the results obtained in terms of specificity from the two models. The Standard Scalar + FA + RF technique achieved perfect specificity of 100%, indicating it correctly identified all true negative cases without any false positives. In comparison, the Standard Scalar + RF method demonstrated slightly lower performance with a specificity of 99.67%, making the feature selection model the more reliable option for ruling out Type II diabetes.

Fig 13. A comparative analysis of the specificity of the two Random Forest Modeling Approaches for Type II Diabetes

The pictorial illustration in Figure 14 showed that the Standard Scalar + FA + RF model demonstrates excellent sensitivity at 98.96%, meaning it successfully identifies nearly all individuals who actually have Type II diabetes. In stark contrast, the Standard Scalar + RF technique shows significantly lower sensitivity at 69.15%, failing to detect the condition in approximately 30% of true cases and making it substantially less effective for reliable screening.

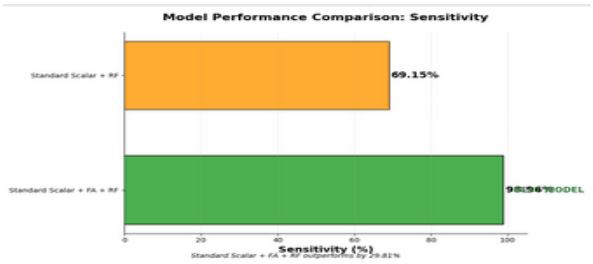

Fig14. A comparative analysis of the sensitivity of the two Random Forest Modeling Approaches for Type II Diabetes

Fig 15 and 16 present the AUC score and the ROC curve for the model with standard scalar technique + FA + RF and the technique with only standard scalar technique + RF. The model using Standard Scalar, Feature selection, and Random Forest (AUC 96.90) performed better than the model using only Standard Scalar and Random Forest (AUC 95.21).

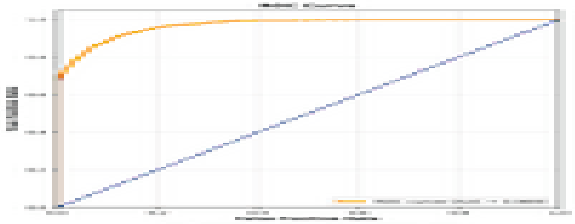

Fig 15. AUC Score for the standard scalar technique + FA + RF

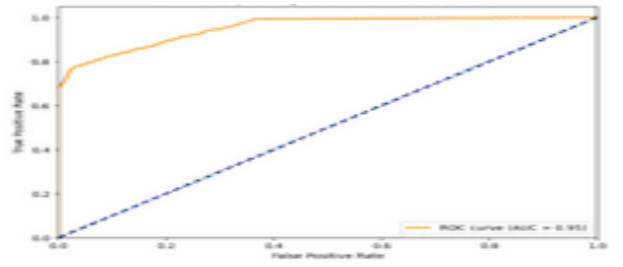

Fig 16. AUC Score for the standard scalar technique + RF

V. DISCUSSIONS

The hybrid Firefly Algorithm and Random Forest model greatly improves diabetes prediction. It achieved a critical sensitivity rate of 98.96%, drastically reducing missed diagnoses. The Firefly Algorithm selects only the most important clinical features for the model. This process removes irrelevant data, which improves the model's accuracy and reliability. The chosen features, like HbA1c and BMI, are well-known medical risk factors. Using familiar factors makes the model's reasoning clear and trustworthy for doctors. Although the algorithm adds an initial step, the final model is faster and simpler to run. The overall model performs better than other common methods like KNN. A current limitation is that it was tested on only one dataset. Future work requires testing on diverse data and creating software for doctors to use.

CONCLUSION

This research developed a highly accurate model for predicting Type II diabetes by optimizing a Random Forest

classifier with the Firefly Algorithm for feature selection. The method identified six key clinical features smoking history, hypertension, BMI, HbA1c, and blood glucose—resulting in a model with exceptional performance: 99.91% accuracy, 99.95% precision, 100% specificity, and 98.96% sensitivity, while also reducing training time. To translate this into practice, the study recommends piloting a Clinical Decision Support System integrated into Electronic Health Records to flag high-risk patients in primary care, alongside ensuring data privacy and creating clinical guidelines.

Future work should focus on validating the model across diverse populations, enhancing explainability using techniques like SHAP, extending research to predict diabetic complications, and conducting cost-effectiveness analyses. These steps aim to transform the model into a trusted, practical tool for early diabetes detection and improved patient outcomes.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium, 2021.

[2] S. Afolabi, N. Ajadi, A. Jimoh, & I. Adenekan, “Predicting diabetes using supervised machine learning algorithms on E-health records,” Informatics and Health, vol. 2, no. 1, pp. 9–16, 2025. [Online]. Available: https://doi.org/10.1016/j.infoh.2024.12.002

[3] M. A. Khan, M. R. Islam, and M. Kabir, “Predictive analytics in diabetes management: A machine learning approach,” Healthcare Technology Letters, vol. 10, no. 3, pp. 45–58, 2023.

[4] R. Patel and J. Smith, “Challenges in traditional diabetes diagnosis and the role of AI,” Journal of Medical Systems, vol. 46, no. 4, pp. 1–15, 2022.

[5] Y. Zhang, B. Liu, and T. Wu, “Machine learning in diabetes prediction: A systematic review,” Artificial Intelligence Review, vol. 56, no. 2, pp. 201–220, 2023.

[6] S. O. Abdulsalam, R. A. Ayofe, M. F. Edafeajiroke, J. F. Ajao, and R. S. Babatunde, “Development of an intrusion detection system using mayfly feature selection and artificial neural network algorithms,” LAUTECH Journal of Engineering and Technology, vol. 8, no. 2, pp. 148–160, Jun. 2024, doi: 10.36108/laujet/4202.81.0241.

[7] S. Gupta, P. Kumar, and A. Jain, “Feature selection techniques for enhanced machine learning models,” IEEE Access, vol. 12, pp. 3456–3470, 2024.

[8] X. S. Yang, “Firefly algorithms for multimodal optimization,” in Stochastic Algorithms: Foundations and Applications. Springer, 2009, pp. 169–178.

[9] L. Wang, H. Zhang, and Q. Zhao, “Nature-inspired algorithms for feature selection in biomedical data,” Computers in Biology and Medicine, vol. 155, p. 106642, 2023.

[10] I. Aljarah, H. Faris, and S. Mirjalili, “Optimizing connection weights in neural networks using the firefly algorithm,” Neural Computing and Applications, vol. 30, no. 7, pp. 2205–2216, 2018.

[11] N. Sharma, P. Tiwari, and V. Mishra, “Metaheuristic algorithms for healthcare data mining: A systematic review,” Computational Biology and Chemistry, vol. 108, p. 107981, 2024.

[12] A. Ogunleye, T. Adekunle, and O. Oluwafemi, “Firefly algorithm-based optimization in healthcare data analysis,” Expert Systems with Applications, vol. 210, p. 118456, 2025.

[13] M. Abdullahi, A. Haruna, and A. Ibrahim, “Metaheuristic optimization in medical diagnostics: A review,” Journal of Healthcare Informatics, vol. 15, no. 2, pp. 112–125, 2023.

[14] S. Rahman and M. Hossain, “Machine learning in diabetes research: Current trends and future directions,” Artificial Intelligence in Medicine, vol. 112, p. 102345, 2024.

[15] A. Ogunleye, T. Adekunle, and O. Oluwafemi, “Firefly algorithm-based optimization in healthcare data analysis,” Expert Systems with Applications, vol. 210, p. 118456, 2025.

[16] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, “Predicting diabetes using supervised machine learning algorithms on E-health records,” Informatics and Health, vol. 2, no. 1, pp. 9–16, 2025. [Online]. Available: https://doi.org/10.1016/j.infoh.2024.12.002

[17] S. K. S. Modak and V. K. Jha, “Diabetes prediction model using machine learning techniques,” Multimedia Tools and Applications, vol. 83, no. 13, pp. 38523–38549, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-16745-4

[18] A. A. Alhussan, N. Al-Dosari, and R. A. Alzaheb, “Chronic complications of diabetes mellitus: A systematic review,” Saudi Medical Journal, vol. 44, no. 3, pp. 221–230, 2023.

[19] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, “A novel classification-based feature selection for early diabetes prediction enhanced with a metaheuristic,” IEEE Engineering in Medicine and Biology Society, vol. 9, pp. 7869–7883, 2021.

[20] S. K. Mohiddin, H. Kousar, P. Sharon, V. S. Krishna, and S. Anupriya, “An approach for early prediction of diabetes using firefly optimization algorithm,” International Journal of Food and Nutritional Sciences, vol. 11, no. 12, pp. 1718–1727, 2022.

[21] N. Aslam et al., “Predicting diabetic patient hospital readmission using optimized random forest and firefly evolutionary algorithm,” International Journal on Advanced Science, Engineering and Information Technology, vol. 11, no. 5, 2021.

[22] N. N. Arslan and D. Özdemir, “A comparison of traditional and state-of-the-art machine learning algorithms for type 2 diabetes prediction,” DergiPark (Istanbul University), 2023. [Online]. Available: https://dergipark.org.tr/tr/pub/jsrc/issue/84404/1365958

[23] C. Liu et al., “Comparison of multiple linear regression and machine learning methods in predicting cognitive function in older Chinese type 2 diabetes patients,” BMC Neurology, vol. 24, no. 1, 2024. [Online]. Available: https://doi.org/10.1186/s12883-023-03507-w

[24] R. Ghabousian, Y. Farhang, K. Majidzadeh, and A. B. Sangar, “Hybrid of particle swarm optimization algorithm and fuzzy system for diabetes diagnosis,” International Journal of Nonlinear Analysis and Applications, in press, 2022. [Online]. Available: http://dx.doi.org/10.22075/ijnaa.2022.29575.4196

[25] K. VijiyaKumar, “Random forest algorithm for the prediction,” in Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[26] C. Bénard, S. Da Veiga, and E. Scornet, “Random forests for survival, regression, and classification (RF-SRC),” The R Journal, vol. 15, no. 2, pp. 1–21, 2023. [Online]. Available: https://doi.org/10.32614/RJ-2023-031

[27] M. Z. Naser and A. Z. Naser, “The firefighter algorithm for optimization problems,” Neural Computing and Applications, 2025. [Online]. Available: https://doi.org/10.1007/s00521-025-11074-z

[28] M. F. Edafeajiroke, S. O. Abdulsalam, M. U. Shuaib, and R. S. Babatunde, “Development of an Intrusion Detection System using ANOVA Feature Selection and Support Vector Machine Algorithms,” Deleted Journal, vol. 4, no. 1, p. 8908, 2024. [Online]. Available: https://doi.org/10.17492/computology.v4i1.2406