<Conference abbreviation>
<Conference Series name> <Volume number and Year> <DOI Number>

Optimizing Diabetes Prediction through Machine Learning and Feature Selection on Health Indicators

Marwa M. Eid[1], Safaa Zaman[2], Amal H. Alharbi[3], Amel Ali Alhussan[4], Doaa Sami Khafaga[5], El-Sayed M. El-kenawy[6,7]

[1]Faculty of Artificial Intelligence Delta University for Science & Technology Mansoura 35111, Egypt Jadara University Research Center Jadara University, Jordan mmm@ieee.org

[2]College of life sciences , Information sciences department Kuwait University S.3zaman@ku.edu.kw

[3]Department of Computer Sciences College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University Riyadh, Saudi Arabia ahalharbi@pnu.edu.sa

[4]Department of Computer Sciences College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University Riyadh, Saudi Arabia aalhussan@pnu.edu.sa

[5]Department of Computer Sciences College of Computer & Information Sciences

Princess Nourah bint Abdulrahman University Riyadh, Saudi Arabia dskhafga@pnu.edu.sa

[6]Delta Higher Institute of Engineering and Technology Department for Communications and Electronics Mansoura 35511, Egypt

[7]Applied Science Research Center Applied Science Private University Amman, Jordan, skenawy@ieee.org

Abstract.

Diabetes mellitus is a chronic metabolic disorder with an increasing global prevalence, necessitating improved diagnostic tools for early detection and management. This study investigates the application of six binary feature selection al- gorithms combined with various machine learning classifiers to predict diabetes in Pima Indian female patients over 21 years old, using health indicators such as glucose levels, BMI, and age. The research systematically compares binary Waterwheel Plant Algorithm (bWWPA), binary Particle Swarm Optimization (bPSO), binary Whale Algorithm Optimization (bWAO), binary Grey Wolf Optimization (bGWO), binary Firefly Algorithm (bFA), and binary Genetic Algorithm (bGA) for optimal feature selection. Subsequently, Random Forest, K-Nearest Neighbors, Decision Tree, and Logistic Regression classifiers were evaluated for diabetes classification performance. The binary Waterwheel Plant Algorithm demonstrated superior feature selection capa- bilities with the lowest average error rate of 0.44554 and highest best fitness score of 0.41054. Among the machine learning models, Random Forest achieved the highest overall performance with 95.66% accuracy, 94.11% sensitivity, 96.82% specificity, and 94.89% F1-score. The results confirm glucose levels, BMI, and age as the most significant predictors for diabetes diagnosis in this population.

Keywords. Diabetes Prediction, Feature Selection, Machine Learning, Random Forest, Pima Indian Dataset, Health Indica- tors.

1. INTRODUCTION

Traditional diagnostic tools such as fasting plasma glucose, oral glucose tolerance tests, and HbA1c remain essential but capture only a limited picture of disease risk, often neglecting socio-demographic and lifestyle factors. ML mod- els have achieved notable predictive success, with ensemble methods such as random forests and gradient boosting often attaining AUC scores above 0.85. Nonetheless, the inclusion of redundant or irrelevant features threatens interpretability and risks overfitting, necessitating effective feature selection frameworks that balance parsimony and predictive perfor- mance [1] [2].

In this study, six bio-inspired binary optimization algo- rithms are investigated as feature selection mechanisms: Bi- nary Waterwheel Plant Algorithm (bWWPA), Binary Particle Swarm Optimization (bPSO), Binary Whale Optimization Al- gorithm (bWOA), Binary Grey Wolf Optimization (bGWO), Binary Firefly Algorithm (bFA), and Binary Genetic Algo- rithm (bGA). These algorithms, inspired by natural, physical, and sociobehavioral systems, provide population-based search strategies capable of addressing combinatorial explosion and avoiding local optima [3]. Their selected feature subsets are evaluated using four interpretable classifiers—Random For- est (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), and Logistic Regression (LR)—with performance measured through stratified k-fold cross-validation using accuracy, AUC, precision, recall, F1-score, and Matthews correlation coeffi- cient (MCC) [4]. Additionally, Shapley Additive Explanations (SHAP) are applied to interpret the contribution of individual features, with plasma glucose and BMI expected as dominant predictors, while other factors such as skin-fold thickness may also yield valuable clinical insights [5].

2. LITERATURE REVIEW

Diabetes is a growing epidemic with severe consequences for patients and healthcare systems, particularly in Taiwan where prevalence continues to rise [6]. A study on 15,000 outpatients in Taipei tested several ML models, showing that the two-class boosted decision tree achieved the best predictive value with an AUC of 0.991. Diabetes mellitus causes hyperglycemia due to insulin dys- function, leading to complications if uncontrolled [7]. Early diagnosis is vital, yet current computational models require im- provement. Studies combining datasets such as PIMA Indian and Medical City Hospital applied feature selection (Spearman correlation), missing value handling (Polynomial Regression), and classifiers (RF, SVM, DNN). Results indicated that a twice-growth deep neural network achieved superior precision and accuracy. Diabetic retinopathy (DR), a common compli- cation, can cause blindness; recent deep learning systems like DeepDR showed high performance in lesion detection and DR grading [8], with AUC values above 0.94. DeepDR-LLM further integrated language models and imaging tools to assist primary care physicians, improving diagnostic efficiency and patient self-care [9].

3. DISCUSSION AND RESULTS

A. Dataset

The data used in this study are derived from the National Institute of Diabetes and Digestive and Kidney Diseases specifically diagnosing indicator for diabetes among the pa- tients. The data in particular refers to female patients only over 21 years of age with Pima Indian ancestry. Included in the dataset is several health metrics that act as predictors for diabetes

presence. Box plots of each of the health measures in the diabetes prediction dataset are shown in Figure 1. These box plots show max, min, second quartile, and first quartile of each feature to compare the differences between diabetic and non-diabetic instances. For example, glucose levels, BMI and age have higher median and upper quartile among diabetics thereby stressing their importance in diabetes prediction. This plot gives the first impression of the main contributing factors of diabetes by just observing for extreme values and scattering of the points.

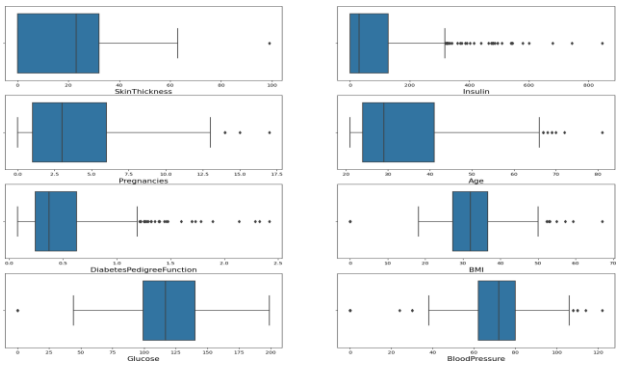

Fig. 1. Box Plots of Key Health Indicators for Diabetes Prediction Dataset

B. Feature Selection Results

The effectiveness of six feature selection techniques namely bWWPA, bPSO, bWAO, bGWO, bFA and bGA is depicted in Table I. Since the performance depends on various factors, performance is evaluated in terms of average error, size of the select, and the fitness values. For instance, the supports of bWWPA were the lowest average of error and the highest best fitness, pointing out that bWWPA has the best performance in selecting important features among all the examined models. Lower error rates and high fitness values suggest that each of the methods utilized is capable of selecting only the most relevant features for diabetes prediction.

C. Machine Learning Results

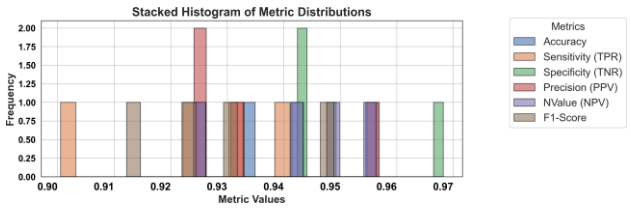

Fig. 2. Stacked Histogram of Diabetes Prediction Model Performance Metrics

The results of different machine learning classifiers, such as Random Forest, KNN, Decision Tree, and Logistic Regression. All predictor models' evaluations of accuracy, sensitivity, specificity and others are given, while the Random Forest Classifier showed the highest accuracy and specificity of all the models tested. Figure 2 presents a stacked histogram detailing the distri- bution of key evaluation metrics for a Diabetes Prediction model leveraging machine learning on health indicators. The graph juxtaposes the frequency of

different metric scores, providing a visual comparison of the model's effectiveness. Precision (PPV), representing the proportion of correctly predicted positive cases among all predicted positive cases, exhibits the highest frequency, peaking at a specific metric value. The distribution of Precision shows frequency of metric at 120.

4. CONCLUSION

This study demonstrates the strong potential of machine learning models, particularly when combined with effective feature selection, in improving the diagnosis of diabetes among Pima Indian women. The findings underscore the importance of combining robust feature selection techniques with strong classifiers to address complex diagnostic challenges. Such systems can provide healthcare practitioners with valuable decision-support tools for identifying high-risk individuals earlier, enabling timely intervention and better allocation of resources. While this research focused on the Pima Indian cohort, the methodologi- cal framework holds promise for broader application across diverse populations and healthcare settings.

REFERENCES

[1] El-Sayed M Elkenawy, Amel Ali Alhussan, Marwa M Eid, and Ab- delhameed Ibrahim. Rainfall classification and forecasting based on a novel voting adaptive dynamic optimization algorithm. Frontiers in Environmental Science, 12:1417664, 2024.

[2] R. Kumar, Vevekanandam, and R. Dey, "IoT-enabled bioimpedance and thermal imaging system for non-destructive fruit maturity assessment," International Journal of Vegetable Science, pp. 1–23, 2025, doi: 10.1080/19315260.2025.2574898.

[3] El-Sayed M El-Kenawy, Abdelhameed Ibrahim, Amel Ali Alhussan, Doaa Sami Khafaga, Ayman EM Ahmed, and Marwa M Eid. Smart city electricity load forecasting using greylag goose optimization-enhanced time series analysis. Arabian Journal for Science and Engineering, pages 1–19, 2025.

[4] R. Kumar, M. K. Singla, S. A. Muhammed Ali et al., "Parameter estimation of proton exchange membrane fuel cell using hybrid grouping biogeography optimization algorithm," Ionics, 2025, doi: 10.1007/s11581-025-06600-x.

[5] Ahmed El-Sayed Saqr, Mohamed S Saraya, and El-Sayed M El- Kenawy. Enhancing co2 emissions prediction for electric vehicles using greylag goose optimization and machine learning. Scientific Reports, 15(1):16612, 2025.

[6] Y. K. Shejwal, A. J. Hati, J. U. Kidav, A. K. Singh, and R. Kumar, "Neural network-based classification of beamforming matrices," International Journal of Satellite Communications and Networking, 2025, doi: 10.1002/sat.70002.

[7] C. C. Olisah, L. Smith, and M. Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220:106773, 2022.

[8] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications, 12(1):3242, 2021.

[9] R. K. Ratnesh, R. K. Chauhan, R. Kumar, M. Afroz, and J. Singh, "Recent advancements of plasmonic sensor technologies in healthcare and electronics," Journal of Molecular Structure, 2025, Art. no. 142945, doi: 10.1016/j.molstruc.2025.142945.