Interpretable Diabetes Risk Prediction using SHAP-Guided Stacking Ensemble Model

Maria Abbas[1], Jamshaid Iqbal Janjua[2,*], Muhammad Zafar Iqbal Karmani[1], Muhammad Zubair Khan[1], Muhammad Bilal Shoaib Khan[1]

[*] *Corresponding Author: jamshaid.janjua@kics.edu.pk*

[1] *Department of Computer Science*
Green International University
Lahore, Pakistan

[2] *Al-Khawarizimi Institute of Computer Science*
University of Engineering & Technology
Lahore, Pakistan

***Abstract*—Diabetes is a significant health issue facing the world today and its early and precise identification is necessary to avoid serious complications. In this study, we present a SHAP-driven stacking ensemble predictive control framework of diabetes risk prediction, which combines the use of Random Forest, Extra Trees, and Logistic Regression as the base models with an XGBoost meta-learner. Borderline-SMOTE is implemented to overcome the issue of class imbalance, which guarantees strong training. The stacking ensemble has the best accuracy of 91.5%, compared to the best base model, Extra Trees (88.9%). SHAP (Shapley Additive exPlanations) interpretability identifies the most important clinical and demographic characteristics, and these offer understandable information about what the model does. The framework is reliable as it is evaluated comprehensively in terms of accuracy, precision, recall, F1-score, specificity, ROC-AUC, and confusion matrices. The suggested methodology has good predictive capability and understandable outcomes, and it will provide a useful instrument to help medical workers make wise choices related to managing diabetes.**

***Keywords*—Ensemble Learning, Explainable Artificial Intelligence, Healthcare Disease Prediction**

I. INTRODUCTION

Diabetes is a persistent metabolic disease, associated with high blood glucose, which has millions of victims across the globe, and is a serious health problem, leading to cardiovascular issues, neuropathy, and renal failure [1], [2]. The identification of people at risk also requires the timely prevention, intervention, and management. Conventional methods of diagnostics are based on clinical evaluation and laboratory studies, which, though effective, are time-consuming, costly, and in the case of a comprehensive screening of large populations are not possible [3]. As a result, efforts have become interested in applying machine learning (ML) methods to predictive diabetes risk modeling, which is faster, less expensive, and more scalable [4], [5].

The research by other scholars has investigated different ML algorithms, such as decision trees, random forests, support machine, and gradient boosting models, to predict diabetes using clinical and demographic characteristics. Although these methods have been shown to perform well in predictive performance, they are not easily interpretable, which in turn restricts the clinical adoption. The medical practitioners need clear models which cannot only give correct predictions, but also clarify the factors behind the decisions. Model interpretability guarantees credibility, responsibility, and insights to action particularly in critical health situations [6].

Ensemble learning methods, including stacking, have been adopted to overcome the two issues: predictive performance and interpretability [7], [8]. In stacking ensembles, several base models are stacked together to enhance the generalization capabilities of the models by capitalizing on the strengths of the other models despite the lack of strong performance by one individual model [9], [10] . In a meta-learner, the predictions are combined together to provide strong results. Nonetheless, stacking models are commonly known as black-box and do not have clear ways of explanation. By incorporating explainable AI (XAI), including SHapley Additive exPlanations (SHAP), it is possible to interpret features at the level of detail to identify the most important factors that drove a particular prediction [11], [12]. This integration is what would make sure that in addition to being able to rely on the predictions provided by the model, the clinicians will be able to comprehend them and justify them as applied to the health of the patient.

Here we present a SHAP-directed stacking ensemble model of predicting diabetes risk. Random Forest [13], Extra Trees [14], and Logistic Regression [15] combined with an XGBoost [16] meta-learner are used as base models. Borderline-SMOTE is used to eliminate imbalance in the classes and to improve the robustness of the model. The framework can predict the outcomes better than single base models and can be explained using SHAP analysis. This framework is an effective instrument that can be used to assist healthcare decision-making due to the combination of performance and transparency. Key contributions of this study:

- Suggested a strong stacking model that has a combination of Random Forest, Extra tree and Logistic Regression as base models and an XGBoost meta-learner which would have better predictive capabilities.
- Applied Borderline-SMOTE to balance the dataset well and enhance the model reliability on minority classes.

- Added SHAP to give feature-level interpretability, which enables clinicians to see the effect of individual input variables on predictions.
- Stacking ensemble was able to achieve high overall accuracy (91.5), which is better than the baseline models (best base: 88.9).
- Provided a clear, practical model that can fit into the context of the actual healthcare decision support in assessing the risk of diabetes.

II. Related Work

Most recent studies in diabetes risk prediction have been pushing more towards the use of machine learning (ML) and ensemble methods to achieve better predictive performance and interpretability. In their study, author employed the use of ensemble ML models using a cohort of the Qatar Biobank that included both clinical and DXA bone health indicators and obtained a maximum accuracy of 87.2% when using feature selection and SHAP interpretation of model outcomes, proving that explainable ensemble models could potentially be utilized in clinical settings [17] On the Pima Indians Diabetes Database, researcher evaluated several classical ML models, such as Random Forest, SVM, Logistic Regression, and XGBoost, in predicting diabetes. They had a moderate accuracy of about 82-83, and SHAP was used to underscore the predictors as influential (glucose and BMI) and the need to focus on explainability in diabetes diagnostics is further justified [18].

Recent researches in related areas of application support important design decisions in our SHAP-based ensemble guide to risk prediction of diabetes. Topology-conscious load balancing Work has shown how adaptation of learning and decision logic to underlying system structure can be used to enhance robustness and efficiency, which is analogous to our use of a meta-learner to integrate complementary base models to be able to generalize better in clinical risk prediction [19]. Similarly, analyses of smart methods of short-term load prediction point out the usefulness of systematically benchmarking a variety of models and reporting overall performance measures in line with our multi-metric validation (accuracy, precision, recall, F1-score, specificity, ROC-AUC, and confusion matrix) to establish a reliable diabetes screening output [20]. The success of ensemble-style modeling on these nonlinear and complex patterns in real-world data is further supported by machine-learning-based residential load prediction, which conceptually resembles our approach of stacking multiple learners to improve predictive performance on heterogeneous clinical and demographic variables [21]. Lastly, AIoT based disease forecaster demonstrates how data-driven predictive frameworks can be generalized to early disease prediction and decision support, which fits our objective of generating reliable but interpretable estimates of diabetes risk supporting actionable health care interventions [22].

Recent works integrate SMOTE, tree-ensemble models and XAI (e.g. SHAP) to provide reliable and explainable healthcare predictions, but tend to test single ensembles instead of stacking. There are other studies that compare several classifiers to SHAP, but extensive feature selection and hybrid oversampling can mask the individual contribution of each algorithm [23]. SHAP is applied in large survey-based analyses but often based on more simple models (e.g. logistic regression and simple trees) rather than stacked ensembles [24]. Comprehensively, as observed in previous reviews, SVM, RF, and deep learning demonstrate promising outcomes, but it has been challenging to achieve a balance between predictive accuracy and clinically significant interpretability across various cohorts [25].

Majority of the research done is based on single or simple ensembles and hardly any attempts have been made to stack different models. Additionally, although SHAP has been applied in explainability, little research has been done to combine SHAP with multi model stacking ensembles with the aim of improving accuracy and facilitating a clear interpretation. Also, there are limited comparisons between such frameworks using multiple base models and balanced evaluation measures, and thus there is a gap in stronger and understandable ensemble frameworks when using diabetes risk prediction.

III. Proposed Methodology

The Fig. 1 represents the general structure of the workflow of the planned diabetes risk prediction framework. The methodology starts with the acquisition of data and then proceeds to data preprocessing, which involves missing value treatment, data normalization, and balancing the classes using SMOTE. Various base learners are trained and stacked together with XGBoost serving as the meta-learner to boost the performance in prediction. The standard classification measures are used to assess the proposed model and the model interpretability is provided by the SHAP-based global and local explanations.

A. Dataset Description

This research employs a publicly available dataset of diabetes that is received in Mendeley Data, containing 5,288 patient records characterized by 14 clinical and demographic variables. The attributes consist of age, gender, body mass index (BMI), pulse rate, systolic and diastolic blood pressure, level of glucose and markers of family history connected with diabetes and cardiovascular diseases. The target variable is the binary status of diabetes, which makes it supervised classification.

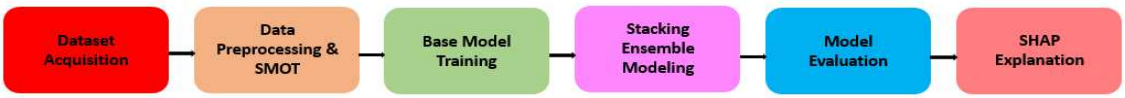

Fig.1 Flow Chart of the Proposed Framework

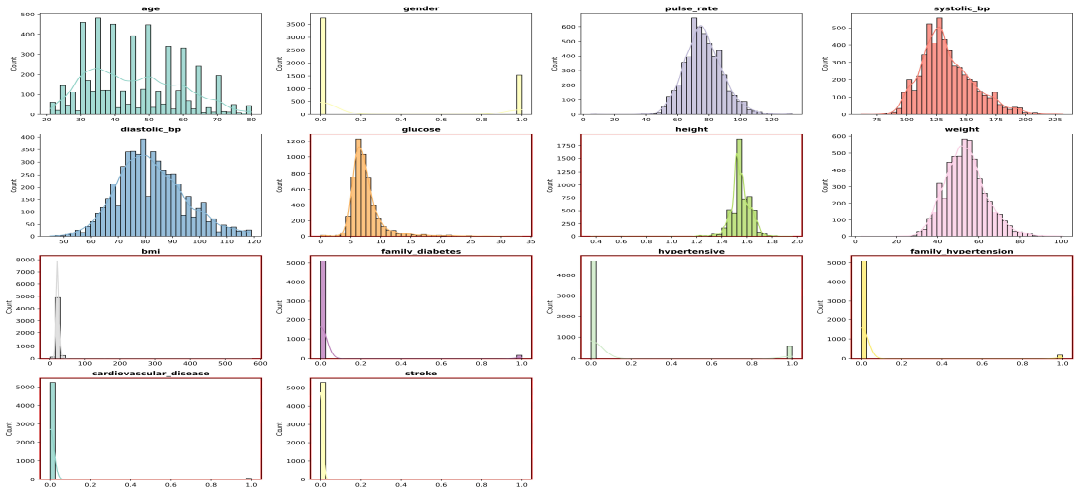

Fig.2 Histogram of Numerical Features

The dataset represents the clinical variability which exists in the real world thus; it is the right dataset to be used in assessing the predictive power and interpretability of the machine learning models in diabetes risk assessment. The Fig. 2 presents histograms of several medical numerical variables (age, gender, pulse rate, blood pressure, glucose, height, weight, etc.) in order to have an idea about their distributions. The small plots indicate each feature and show the frequency of various ranges of values, which allows observing the location of the data concentration and the balance or occurrence of skewness. This value is a correlation heatmap that indicates the strength of the relationship between every pair of features of the dataset. A high positive correlation is represented by darker red cells, a high negative correlation is represented by darker blue colors and values near zero (light colors) indicate significant or no linear relationship between the variables as shown in Fig. 3.

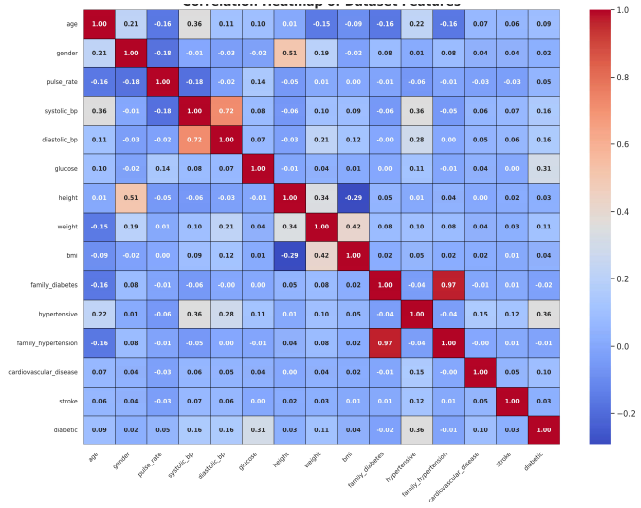

Fig.3 Correlation Heatmap of Dataset

B. *Data Preprocessing and Balancing*

The Before training a model, an extensive preprocessing pipeline is used to guarantee data quality and stable learning will be used. First, the data is analyzed regarding missing and incongruent values. Given that the features are quantitative and presented in written records, entries with unfinished or invalid data are handled attentively to maintain data integrity. The input features are made to a standard scale whereby they are represented to a uniform scale to avoid being biased in favor of the attribute with a larger numerical range especially in optimization based learning algorithms.

Borderline Synthetic Minority Over-Sampling Technique (Borderline-SMOTE) is used to solve the problem of class distribution imbalance with the training data alone only. Under this approach, synthetic samples are created close to the decision border allowing the models to learn more discriminative patterns between diabetic and non-diabetic cases and lessening the chances of overfitting. To guarantee the information leakage does not occur when evaluating the model, it is sufficient to apply data balancing to the training set. Lastly, stratified train test splitting is applied to the preprocessed data splitting them in proportions equal to that of the original classes. Such a preprocessing plan preconditions a high-quality experimental base, which provides the opportunity to make a fair comparison of models and assess their predictive performance and interpretability reliably.

C. *Machine Learning Models*

In order to enhance the prediction accuracy of diabetic and non-diabetic classification, we use a stacking ensemble technique. The ensemble makes use of various base models that determine different patterns in the data, where a meta-model is used to combine the predictions in the best way.

D. *Random Forest (RF)*

Random Forest represents non-linear interaction between clinical characteristics. Its prediction capacity to combine the forecasting of a set of decision trees offers to minimize overfitting issues on this dataset and remains responsive to the slightest changes in the blood parameter characteristics.

E. Extra Tree (ET)

Extra Trees provides an extra amount of randomization to tree building, which is especially useful in learning latent interdependences between features which may not be fully utilized by RF. This is applicable in our dataset whereby some feature combinations would play a larger role in the prediction.

F. Logistic Regression (LR)

Logistic Regression creates a linear view of the relationships between features and targets so that powerful linear relationships in the set of data are not ignored. Its additional stability and interpretability as a tree based model complements the tree-based models.

G. Stacked Ensemble

XGBoost is used as the meta-model to combine the predictions of RF, ET and LR. XGBoost is a system that allows heterogeneous input by the use of numerous base models and optimizes the predictive performance by gradient boosting. XGBoost is able to blend the strength of the base models and recompense the weakness of a single model, thus leading to overall more accuracy and strength in the prediction task.

H. Model Evaluation

To measure the performance of the proposed stacking ensemble, a number of metrics are used in order to have a strong evaluation of the predictive power of the model. F1-score and accuracy measure the total accuracy and a balance between precision and recall of classes diabetic and non-diabetic. Confusion matrices are used to depict right and wrong predictions of each class and this information offers a reflection on the strengths and weaknesses of the model. Also, bar graphs are employed in comparing the performance of the individual base models as well as the compare and contrast performance of the stacking ensemble of these models, showing the level of improvement that is made by combining models together. This assessment framework is comprehensive in the sense that it will guarantee quantitative and visual knowledge of the effectiveness of the model.

IV. Result

This diabetes prediction framework applies base classifiers; among them are Random Forest (RF), Extra Trees (ET), and Logistic Regression (LR), and a stacked ensemble model with XGBoost as the meta-model. All the models are trained and evaluated with Python on the Google Colab platform. The data is divided in terms of training and testing sets in a proportion of 80: 20. The offered solution is effective in terms of dividing diabetic and non-diabetic individuals. Measurements of evaluation used in determining the performance of the two individual base models and the stacked ensemble are accuracy, F1-score, precision, recall and specificity, which are outlined and discussed below.

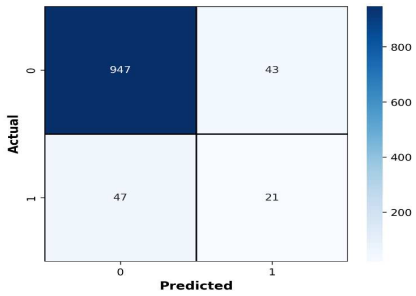

Fig.4 Confusion Matrix of Stacked Model

As shown in Fig. 4, stacking ensemble confusion matrix depicts its performance levels in the accurate classification of diabetic and non-diabetic cases. It indicates the number of true positives, false positives, true negatives and false negatives which reveals the capacity of the model to minimize misclassifications. This analysis shows the predictability and consistency in the prediction between the two classes in the ensemble.

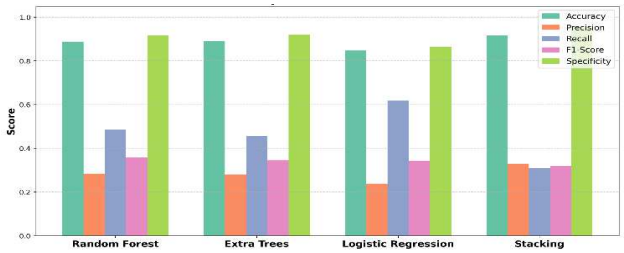

Fig.5 Bar Chart Comparsion across Base vs Stacked Model

To compare the performance of the base models and the stacking ensemble on five evaluation metrics, including accuracy, F1-score, precision, recall, and specificity, the performance is compared with the help of a bar chart. As shown in Fig. 5, the stacking ensemble would always perform better than the base models individually in every measure. The effectiveness of multiple model combination is well illustrated in this comparison as the overall predictive ability is better. The bar chart is an easy method of evaluating and conveying the performance increment that the stacking approach has brought.

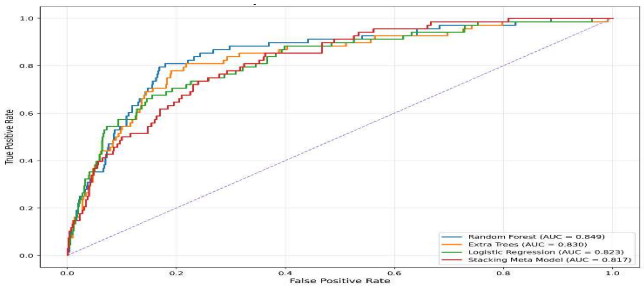

Fig.6 ROC Curve of Base Models and Stacked Model

ROC curves are plots that are used to compare the performance of the base models with the stacking ensemble by displaying the trade-off between sensitivity and false positive rate. The stacking ensemble has the best AUC and it has a better capability of differentiating diabetic and non-

diabetic cases as shown in Fig. 6. This proves that the ensemble is better than individual base models in terms of classification performance on the whole.As shown in Fig. 7, the precision- recall curves are used to compare the trade-off between precision and recall of the base models and the stacking ensemble. The stacking ensemble reaches the largest area under the Precision-Recall curve, which means that it better performs at the task of correctly classifying the cases of diabetes and has the lowest false positives. It shows that the ensemble has the most appropriate balance between precision and recall and it is more accurate at classification when compared to individual base models.

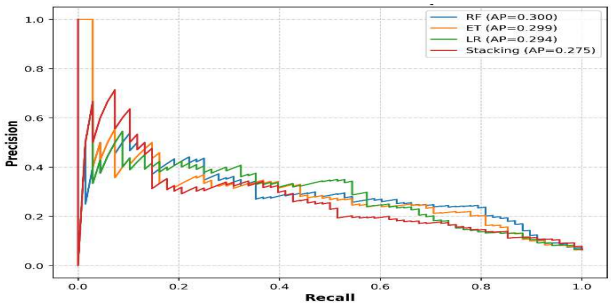

Fig.7 Precision-Recall Curve of Base Models and Stacked Model

The learning curve of the stacking ensemble shows the trends in the accuracy of the training and validation as the training sample size grows as shown in Fig. 8. Based on the curve, it can be seen that the model has always high training accuracy and increasing validation accuracy as more data is provided, which proves effective learning and generalization. This discussion proves that the ensemble is not overfitted, and bigger training sets are advantageous to the predictive performance in discerning diabetic and non-diabetic cases.

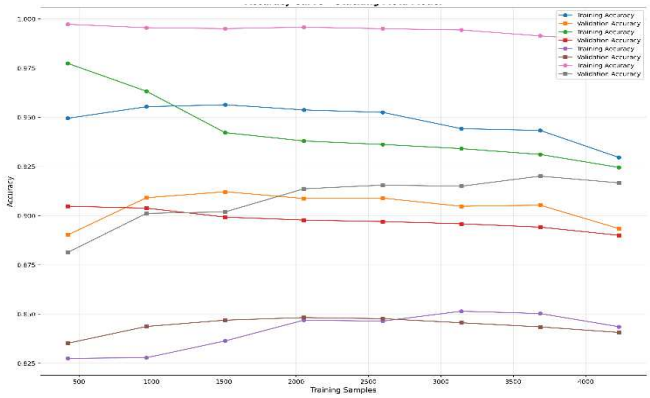

Fig.8 Learning Curve of Stacked Meta Model

To explain the predictions of the base models, SHAP (SHapley Additive exPlanations) is employed to measure how each feature contributes to classifying between diabetic and non-diabetic cases. The SHAP beeswarm plot graphs help visualize the significance of features and the orientation of their effect, whereas the bar charts provide the summary of the features significance in the world expressed by mean absolute SHAP values.

In the case of the Random Forest model, hypertensive status, glucose level and systolic blood pressure were the most influential features that affect model predictions in the analysis as shown in Fig. 9. The increased feature values tend to have a positive effect on the prediction of the cases of diabetes, and it means that clinical risk factors are powerful.

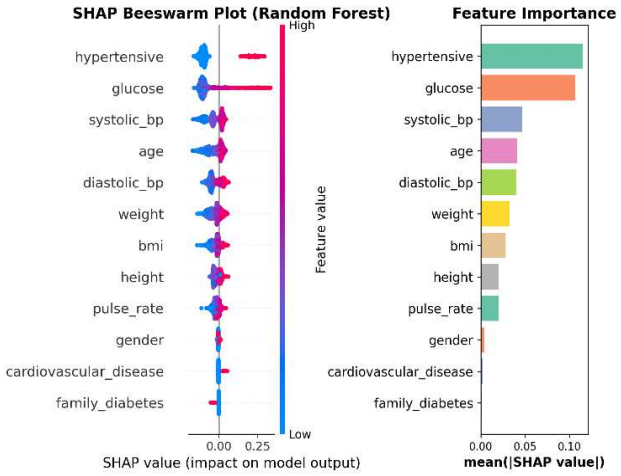

Fig.9 SHAP Interpretation for Random Forest

As shown in Fig. 10, the Extra Trees model identifies hypertension and glucose to be the most significant predictors after which blood pressure-related features and age are considered as the next most important ones. The propagation of SHAP values is an indicator that the model is able to encode the interaction of complex features, which supports its position as a powerful non-linear base learner.

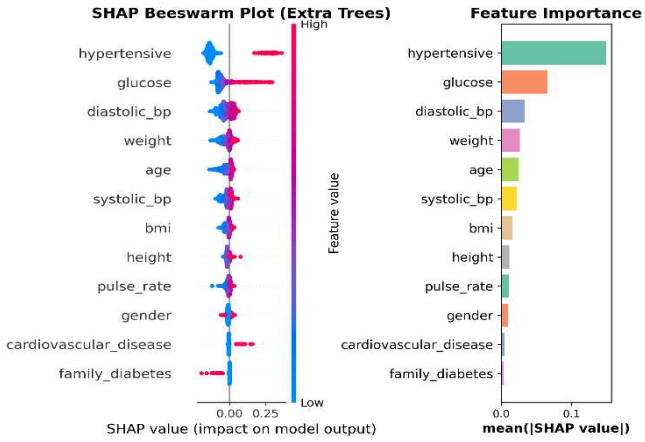

Fig.10 SHAP Interpretation for Extra Trees

In the case of Logistic regression, coefficient values represent the feature importance and they give a linear and interpretable perspective of its impact as shown in Fig. 11. Aspects like hypertensive status, family history, and cardiovascular measures are more associated with predicting diabetes, and the direction of coefficients are used to determine whether these aspects predict the likelihood of diabetic outcome or not.

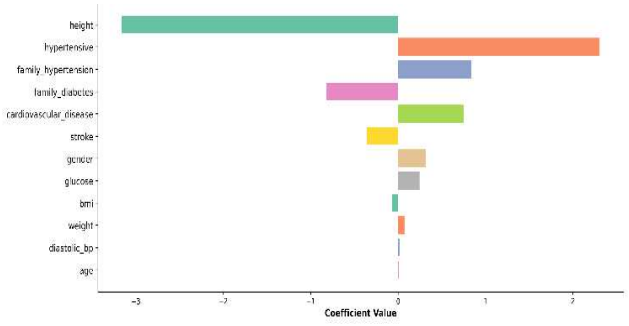

Fig.11 Feature Importance for Logistic Regression

V. Conclusion

This paper introduces a machine learning-driven model of diabetic and non-diabetic individual's classification based on a stacking ensemble algorithm. Base models used are Random Forest, Extra Trees, and Logistic Regression whereas the XGBoost is used as a meta-model to combine their prediction. Extensive testing and explainable AI (XAI) analysis shows that the proposed framework has a good performance in addition to offering meaningful model explainability. In spite of the fact that the proposed approach demonstrates a high level of prediction, it is tested on a single dataset, which can be improved by the validation with larger and more heterogeneous clinical datasets. Future research will concentrate on the multi-center data extension of the framework, as well as further investigation.

References

[1]. Armghan, J. Logeshwaran, S. Sutharshan, K. Aliqab, M. Alsharari, and S. K. Patel, "Design of biosensor for synchronized identification of diabetes using deep learning," Results in Engineering, vol. 20, p. 101382, 2023.

[2]. G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. Meddage, and U. Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," Healthcare analytics, vol. 5, p. 100301, 2024.

[3]. Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," Scientific Reports, vol. 12, no. 1, p. 1170, 2022.

[4]. F. A. Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Materials Today: Proceedings, vol. 80, pp. 3200-3203, 2023.

[5]. 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 and Applications, vol. 83, no. 8, pp. 24153-24185, 2024.

[6]. M. A. Hama Saeed, "Diabetes type 2 classification using machine learning algorithms with up-sampling technique," Journal of Electrical Systems and Information Technology, vol. 10, no. 1, p. 8, 2023.

[7]. Q. Saihood and E. Sonuc, "A practical framework for early detection of diabetes using ensemble machine learning models," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 31, no. 4, pp. 722-738, 2023.

[8]. Dutta, "Early prediction of diabetes using an ensemble of machine learning models," International Journal of Environmental Research and Public Health, vol. 19, no. 19, p. 12378, 2022.

[9]. S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," Frontiers in Genetics, vol. 14, p. 1252159, 2023.

[10]. R. Ganguly and D. Singh, "Explainable artificial intelligence (xai) for the prediction of diabetes management: An ensemble approach," International Journal of Advanced Computer Science and Applications, vol. 14, no. 7, 2023.

[11]. F. Curia, "Explainable and transparency machine learning approach to predict diabetes develop," Health and Technology, vol. 13, no. 5, pp. 769-780, 2023.

[12]. P. B. Khokhar, V. Pentangelo, F. Palomba, and C. Gravino, "Towards transparent and accurate diabetes prediction using machine learning and explainable artificial intelligence," arXiv preprint arXiv:2501.18071, 2025.

[13]. N. Noviyanti and A. Alamsyah, "Early detection of diabetes using Random Forest algorithm," Journal of Information System Exploration and Research, vol. 2, no. 1, 2024.

[14]. L. Malviya, R. Dangi, A. Jadhav, and J. Kishore, "Optimization-Based Hyperparameter Tuning Using Extra Trees to Classify Type-2-Diabetes Mellitus," in 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), 2025: IEEE, pp. 465-470.

[15]. M. R. Belgaum et al., "Enhancing the efficiency of diabetes prediction through training and classification using PCA and LR model," Annals of Emerging Technologies in Computing (AETiC), vol. 7, no. 3, pp. 78-91, 2023.

[16]. Maulana et al., "Machine learning approach for diabetes detection using fine-tuned XGBoost algorithm," Infolitika Journal of Data Science, vol. 1, no. 1, pp. 1-7, 2023.

[17]. Alsadi et al., "An ensemble-based machine learning model for predicting type 2 diabetes and its effect on bone health," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 144, 2024.

[18]. İ. Kırbaş and A. Çifci, "Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset," Balkan Journal of Electrical and Computer Engineering, vol. 13, no. 2, pp. 128-139, 2025.

[19]. T. A. Khan, M. S. Khan, S. Abbas, J. I., S. S. Muhammad and M. Asif, "Topology-Aware Load Balancing in Datacenter Networks," 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), Bandung, Indonesia, 2021, pp. 220-225, doi: 10.1109/APWiMob51111.2021.9435218.

[20]. Ahamed, N. Ahmed, J. I., Z. Hossain, E. Hasan and T. Abbas, "Advances and Evaluation of Intelligent Techniques in Short-Term Load Forecasting," 2024 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2024, pp. 1-9, doi: 10.1109/ICCA62237.2024.10927804.

[21]. W. Alomoush, T. A. Khan, M. Nadeem, J. I., A. Saeed and A. Athar, "Residential Power Load Prediction in Smart Cities using Machine Learning Approaches," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-8, doi: 10.1109/ICBATS54253.2022.9759024.

[22]. T. Abbas, J. I. and M. Irfan, "Proposed Agricultural Internet of Things (AIoT) Based Intelligent System of Disease Forecaster for Agri-Domain," 2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2023, pp. 1-6, doi: 10.1109/ICCA59364.2023.10401794.

[23]. Y. Barzegar, A. Barzegar, F. Bellini, F. D'Ascenzo, I. Gorelova, and P. Pisani, "Machine Learning Pipeline for Early Diabetes Detection: A Comparative Study with Explainable AI," Future Internet, vol. 17, no. 11, p. 513, 2025.

[24]. B. Barman, H. K. Choudhury, and B. Jajodia, "Interpretable machine learning for diabetes risk prediction: a large-scale analysis of Indian national survey data," Discover Public Health, vol. 22, no. 1, p. 832, 2025.

[25]. P. B. Khokhar, C. Gravino, and F. Palomba, "Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review," Artificial intelligence in medicine, p. 103132, 2025.