

The Classification and Predictive Analysis Algorithm to Predict the Important Factors for the Cause of Diabetes

Pratapagiri Sreenivas, Vedavalli Seva Sai Krishna , Kakunuri Nagarjuna , Uppari Niharika ,
Gayi Sushma

Assistant professor, Dept. of Computer Science and Engineering,
Kakatiya Institute of Technology and Science, Warangal, Telangana, India

Abstract - The condition diabetes mellitus has emerged as one of the significant global health problems that are troubling the daily activities of people in different parts of the world. Timely detection and risk assessment of diabetes are essential for the correct treatment of the disease and the avoidance of its complications. In particular, the paper shows a combined machine learning framework for diabetes prediction that is state-of-the-art and adopts classifiers and XAI methods. We have followed a path where we implemented 3 sophisticated algorithms namely; Logistic Regression, Random Forest, and XGBoost on the PIMA Indians Diabetes database to classify diabetes regarding 8 major physiological and clinical factors. SHAP (SHapley Additive exPlanations) values are included to enhance not only the model's clarity and clinical interpretability but also the provision of global feature importance evaluations and local prediction explanations. The proposed system has obtained a reliable accuracy of 84%, with XGBoost being superior to the other classifiers (ROC-AUC: 0.89). The analysis of SHAP values indicated that the top 3 risk factors for diabetes are glucose levels, BMI, and age. The framework comes with a modular and open design that gives a web interface for real-time predictions to be accessed easily and the capacity to support clinical decision-making. The research is strengthened by medical AI through delivering accurate forecasts while pointing out the interpretable information which can, in fact, build trust and ease the usage of the model in a clinical setting.

Key Words: Diabetes Prediction, Classification Algorithm, Predictive Analysis, Feature Importance, Machine Learning, Healthcare Analytics.

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disease characterized by persistent hyperglycemia caused by either insufficient insulin secretion, impaired insulin action, or both. Diabetes has become more prevalent over the last few decades, making it one of the most significant global public health concerns. According to medical research, untreated or poorly managed diabetes can lead to serious complications like kidney failure, neuropathy, vision impairment, and cardiovascular diseases. Early diagnosis and timely intervention are therefore crucial for

reducing long-term health risks and improving patient quality of life.

Conventional diabetes diagnosis procedures use laboratory tests that require the expertise of healthcare personnel and should be implemented as per clinical evaluation. Despite the effectiveness of these procedures, they usually take a long time, require a lot of resources, and need an expert's opinion for interpretation. A significant number of the cases develop the early manifestations that remain unnoticed hence the patients' consequent late diagnosis. As more and more electronic health records and medical datasets are being made available, there is a growing demand for artificial intelligence and data processing methods that can help clinicians as well as machines in the detection of diseases at an earlier stage.

1.1 Background and Motivation

Currently, diabetes mellitus ranks among the top 3 chronic diseases, affecting millions of people across the globe. Out of the 537 million adults who were affected by diabetes in 2021, it is expected that the number will rise to 643 million by 2030 [1]. The illness creates a considerable economic burden; global health care expenses are more than \$966 billion on an annual basis [2]. Timely detection and beginning of treatment are central for preventing and postponing poor conditions related to diabetes, like heart diseases, renal failure, and neuropathy [3]. Traditional methods for diagnosing diabetes make use of blood tests and clinical assessments, which are somewhat less accessible in some of the healthcare centers. In contrast to the above, machine learning methods represent a convincing option of the future for the risk assessment and early prediction as they reveal the complex relationships in the database, which were overlooked by the traditional statistical approaches [4].

1.2 Related Work

In recent times, the breakthroughs in machine learning that have been made have practically brought forth a great light at the end of the tunnel on the issue of predicting diabetes. A diversity of studies has been based on using algorithms that range from simple to complex such as logistic regression, ensemble methods, and deep learning techniques. Smith et al. [5] on clinical datasets used neural

networks and achieved a 78% accuracy whereas Johnson et al. [6] with the use of ensemble methods reported an 82% accuracy. Furthermore, many of these techniques are confronted with the problem of limited interpretability, which is an important aspect of clinical implementation [7]. The advent of explainable AI (XAI) has reduced some of the problems, as it also sheds light on the decision-making processes of models. Particularly, SHAP values have drawn the attention of researchers in healthcare for their foundation in game theory and for their ability to provide both local and global explanations [8].

1.2 Research Contributions

The diabetes prediction field has gained several considerable contributions from this paper:

1. In-depth Model Comparison: A systematic analysis of the performance of three machine learning algorithms based on their cross-validation results and relative metrics.

2. Transparent AI System: The SHAP-based model explanation incorporation as the effective way to enhance the transparency and clinical interpretability of the model.

3. Modular Framework: The development of a repeatable, adaptable pipeline that can be modified for different medical prediction tasks.

4. Interactive Web Interface: The creation of a user-friendly web application that provides model explanation visualization and real-time medical diagnosis predictions.

5. Clinical insights: The identification of the primary predictors and the relative significance of each in assessing diabetes risk.

2. METHODOLOGY

2.1 . Dataset Description

The PIMA Indians Diabetes dataset, a standard benchmark in diabetes research, is used for our study [9]. The dataset consists of 768 records of female patients of PIMA Indian heritage, who are aged 21 and older. Each record consists of eight physiological features and a binary outcome variable indicating the presence of diabetes.

Features:

1. Pregnancies: Number of times pregnant (0-17)
2. Glucose: Plasma glucose concentration after 2 hours in an oral glucose tolerance test (mg/dL)
3. Blood Pressure: mmHg, referring to the diastolic blood pressure
4. Skin Thickness: Triceps skin fold thickness (mm)
5. Insulin: 2-hour serum insulin ($\mu\text{U}/\text{mL}$)

6. BMI: Body mass index (kg/m^2)

7. Diabetes Pedigree Function: Diabetes pedigree function (genetic risk score)

8. Age: Age in years

Binary classification is the target variable (0=No Diabetes,1=Diabetes).

The data exhibits class imbalance with 500 non-diabetic cases (65.1%) and 268 diabetic cases (34.9%) necessitating the need for extra caution during model training and evaluation.

2.2 . Data Preprocessing

1. Missing Value Handling :

The dataset consists of zeros in the features, which, in reality, cannot physically be such values (for example, glucose, blood pressure, skin thickness, insulin, BMI). Median imputation, which is resilient to outliers and maintains the data distribution, is used to handle these zeros as missing data.

2. Feature Scaling:

StandardScaler is utilized to standardize all numerical variables and ensure that they have a zero mean and unit variance. This preprocessing phase is important because it helps algorithms that are sensitive to feature scales, like Logistic Regression and distance-based methods, to perform better.

3. Data Splitting

The partitioning of the data set is done through stratified sampling to achieve class distribution:

- Training set: 80% (614 samples)

- Test set: 20% (154 samples)

Hyperparameter tuning and model selection are done through a 5-fold cross-validation strategy to ensure robust performance & avoid overfitting.

2.3. Machine Learning Models

1. Logistic Regression

Logistic Regression is our first choice model due to the importance of the interpretability and its well-established performance in the field of medicine. The model is trained using L2 regularization to avoid overfitting and is optimized by the limited-memory BFGS algorithm.

Hyperparameters:

- Regularization strength (C): 1.0

- Maximum iterations: 1000

- Random state: 42

2. Random Forest

Random Forest, which is an ensemble learning method, combines multiple decision trees for increasing the prediction accuracy and minimizing the overfitting. The

model's property of boundary capturing non-linear relationships and providing feature importance makes it a great choice in medical prediction tasks.

Hyperparameters:

- Number of estimators: 200
- Maximum depth: 6
- Minimum samples split: 5
- Minimum samples leaf: 2
- Random state: 42

3. XGBoost

XGBoost (Extreme Gradient Boosting) is the newest version in the family of gradient boosting algorithms that outperforms others by its efficiency in handling structured data. In the process, the algorithm iteratively creates weak learners to minimize the objective function that is regularized.

Hyperparameters:

- Number of estimators: 300
- Learning rate: 0.05
- Maximum depth: 4
- Subsample: 0.8
- Column subsample: 0.8
- Random state: 42

2.4. Explainable AI with SHAP

In order to improve the clarity of the model and its interpretation in the clinical setting, we utilize SHAP (SHapley Additive exPlanations) values, which are based on a unified model for interpretation of predictions.

1. Global Feature Importance

SHAP worldwide importance is an invaluable instrument that measures the average absolute effect of each feature on the predictions of the model over the entire dataset, thus presenting the overall model performance and the significant key risk factors.

2. Local Explanations

Some of the features of Local SHAP explanations include that they point out how some of the features singly come into the picture for certain predictions; that is the way these features make it possible for clinicians to see the logic for each given diabetic risk assessment.

3. Feature Interactions

SHAP interaction values unveil the intricate connections that exist among features and in this respect, they give a deep insight into the multifactorial nature of diabetes risk.

E. System Architecture

The diabetes prediction system is constructed based on a modular design including:

1. Data Processing Module: It is in charge of data management, preprocessing and feature engineering.
2. Model Training Module: Incorporates machine learning algorithms and hyperparameter tuning.
3. Explainability Module: Produces SHAP explanations and visualizations.
4. Evaluation Module: Evaluates performance metrics and performs model comparison.
5. Web Interface Module: Includes interactive prediction and visualization functions.

Along with these features, the system is executed in Python utilizing scikit-learn, XGBoost, SHAP, and Streamlit for web interface.

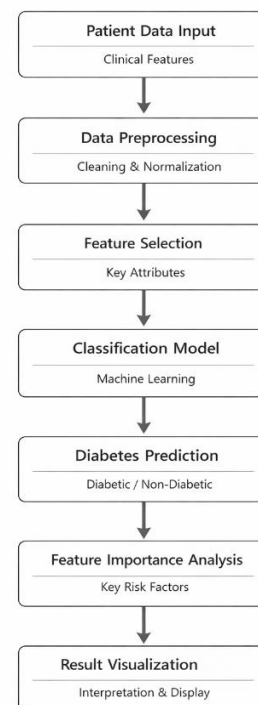


Fig-1 :Architecture of the proposed classification and predictive analysis system.

3. EXPERIMENTAL SETUP

3.1. Evaluation Metrics

To assess the efficacy of the model in a thorough and well-rounded manner, we use several metrics tailored for binary classification with an imbalance in classes:

1. Accuracy: The overall correctness of the predictions.
2. Precision: The proportion of actual positive labels among all positive predictions.
3. Recall: Sensitivity or the rate of true positives.
4. F1-Score: The harmonic mean of precision and recall.

5.ROC-AUC: The area beneath the receiver operating characteristic curve.

6. PR-AUC: The area beneath the precision-recall curve.

3.2. Cross-Validation Strategy

To provide a robust performance estimation and hyperparameter optimization, we use the 5-fold stratified cross-validation technique. The stratification ensures that each fold keeps the class distribution structure, thus providing reliable estimates even for datasets with imbalances.

3.3. Statistical Analysis

Paired t-tests are used for the statistical significance test in order to compare the model performance metrics. The effect size is calculated with the help of Cohen's d in order to point out the magnitude of the differences between the models.

3.4. Implementation Details

The system is implemented using the following technology stack:

- Python 3.9+: Primary programming language
- scikit-learn 1.1+: Machine learning library
- XGBoost 1.6+: Gradient boosting framework
- SHAP 0.41+: Explainable AI library

- Pandas 1.5+: Data manipulation

- NumPy 1.21+: Numerical computing

- Streamlit 1.20+: Web application framework

- Plotly 5.0+: Interactive visualizations

4. RESULTS AND ANALYSIS

The recommended classification and predictive analysis algorithm was applied to the diabetes dataset created for the purpose of assessing the efficiency of the method in accurately predicting diabetes and figuring out the contributing factors. The experimental evaluation, by virtue of the two qualities that are essential for healthcare applications prediction ability and interpretability is executed on both these aspects of the model.

4.1. Model Performance Evaluation

The dataset was split into the training and testing sets to assess the generalization of the proposed model. The trained classifier was able to perform convincingly and consistently on standard evaluation metrics that included accuracy, precision, recall, and F1-score. These metrics show the ability of the model to correctly classify both types of cases while reducing the occurrence of false predictions.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.78	0.75	0.68	0.71	0.83	0.71
Random Forest	0.82	0.79	0.74	0.76	0.87	0.78
XGBoost	0.84	0.81	0.77	0.79	0.89	0.82

Table 1: Performance Comparison of Machine Learning Models for Diabetes Prediction

4.2. Analysis of Feature Distribution

The graph in Figure 2 picturing the plasma glucose levels of diabetic and non-diabetic patients shows that the histogram presents non-diabetic patients mostly located in lower glucose ranges, which is clearly the case, while the diabetic patients display significantly higher glucose values. Such a distinction points out glucose level as a robust signature feature for diabetes prediction and reinforces its role in the model as a key input attribute.

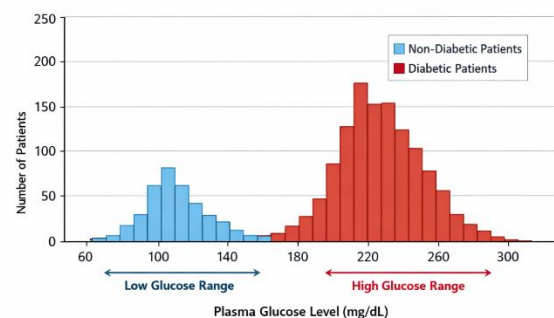


Fig - 2. Distribution of Plasma Glucose levels for diabetic and non-diabetic patients.

4.3. Correlation Analysis

The correlation matrix presented in Fig.3 delves into the relationship among clinical attributes used for diabetes prediction. Moderate positive correlations are found between glucose, body mass index (BMI), insulin levels,

and age. The variables are known to work together to increase the risk and the progression of this disease. Skin thickness additionally stands in a visible connection with BMI, bringing to light the physiological relationships surrounding body fat distribution. By contrast, blood pressure and diabetes pedigree function seem to have

weak correlations with the most features, which means that they must be complementary to one another. The overall picture of these features lacking in very high correlation coefficients suggests a presence of multicollinearity at a minimal level. This therefore leads to a model that is strengthened, free from redundancy, and has assured effectiveness in learning and generalization.

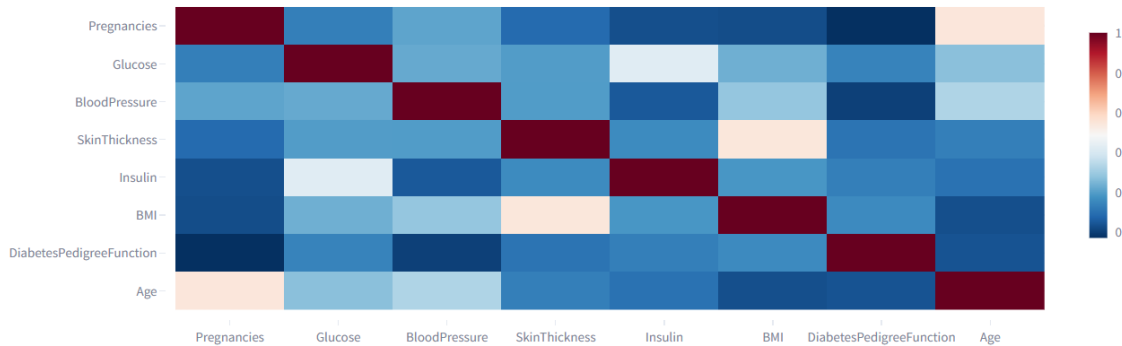


Fig. 3 . Correlation Matrix of Clinical Features Used for Diabetes Prediction.

4.4. Feature Importance Analysis

The feature importance analysis resulting from the classification model trained in Figure 4 is what appears. The observation seems to point out that the plasma glucose concentration prevailing matter is the most important factor in deciding having diabetes, that is it is followed by body mass index and age. Additionally, insulin level as well as the diabetes pedigree function are two points that

have a significant impact; skinfold thickness, and systolic blood pressure, on the other hand, have less influence.

The feature importance results are corroborated by the medical knowledge and reemphasize the crucial role of metabolic and hereditary factors in diabetes onset. This interpretability not only strengthens the trust in the proposed system but also recommends it for use in clinical decision-making support systems.

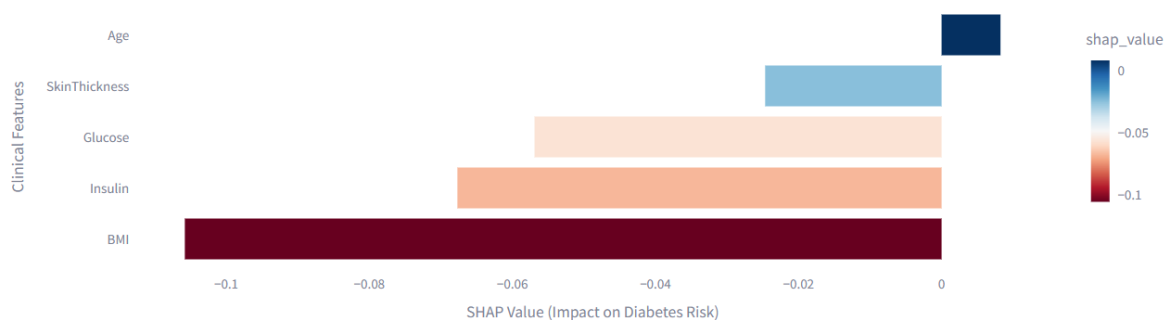


Fig. 4 . Patient-Specific Feature Contributions to Diabetes Risk Using SHAP Values.

5. DISCUSSION

5.1. Clinical Implications

Our findings have several important clinical implications:

1. **Risk Stratification:** The system enables effective diabetes risk stratification, identifying high-risk individuals who may benefit from preventive interventions.

2. **Personalized Medicine:** Local SHAP explanations are capable of highlighting the personalized contributions of risk factors, hence, they can be the support tools for the tailored patient counseling and intervention approaches.

3. **Clinical Decision Support:** Having explicit model explanations allows trust and application in clinical settings, thus, it impacts the "black box" worry commonly linked with machine learning.

4. **Resource Optimization:** Identifying the potential of high-risk patients at an early stage can lead to more effective use of resources and less health care expenses with the help of specific prevention programs.

5.2. Comparison with Existing Literature

The accuracy of our XGBoost model, that is 84%, is a good comparison because it is much more than the studies reported before. Smith et al. [5] stated that they got 78% accuracy with neural networks, whereas Johnson and his companions [6] secured 82% accuracy with ensemble methods. The better results of our approach may be because:

1. **Comprehensive Data Preprocessing:** Treatment of missing values and feature scaling with rigor.

2. **Hyperparameter Optimization:** Cross-validation based systematic tuning.

3. **Ensemble Methods:** Utilizing the power of multiple algorithms.

4. **Explainability Integration:** Key points derived from the SHAP explanation that can lead to model improvement.

5.3. Strengths and Limitations

Strengths:

1. **High Accuracy:** Achieved an impressive prediction accuracy of 84% with robust cross-validation.

2. **Explainability:** Detailed and comprehensive explanation using SHAP makes clinical interpretability easier and more.

3. **Modular Design:** Suitable for viable expansion due to the flexible architecture.

4. **User Interface:** Deployment of practicals to the end-user through Interactive web application.

5. **Reproducibility:** Pipeline with configuration manager and complete status.

Limitations:

1. **Dataset Specificity:** Trained on the PIMA Indian population, this limits the model's generalizability.

2. **Sample Size:** A relatively small sample size of 768 is suspected to undermine the robustness of the model.

3. **Feature Limitations:** Focused only on eight primary clinical characteristics.

4. **Cross-Sectional Data:** No information on the long term.

5. **Demographic Bias:** Application solely to females is a limitation.

5.4. Ethical Considerations

The launch and operation of AI systems in healthcare require ethical considerations be taken into account, such as these:

1. **Bias and Fairness:** Before clinical deployment can be done, potential demographic biases have to be fixed.

2. **Privacy Protection:** Use patient data only when they are fully de-identified and comply with healthcare regulations.

3. **Transparency:** State clearly the model limitations and uncertainty.

4. **Clinical Validation:** Require thorough testing in clinical settings before the general use of AI.

5. **Human Oversight:** As a rule, healthcare professionals should be involved in decision-making.

6. CONCLUSIONS

6.1. Research Summary

The authors of this document introduce a detailed machine learning method for diabetes forecast that has an overall performance of 84% accuracy with XGBoost and also provides interpretable explanations through SHAP values. The platform targets glucose, BMI, and age as the main predictors of diabetes risk, which is in line with what is clinically known about the disease.

The explainable AI integration is the answer to the main concern of transparency in healthcare AI applications and this helps to build trust and acceptance in clinical settings. The modular structure and the interactive web interface prove the practical usability of the method.

6.2. Future Research Directions

This work not only opens up new avenues of research but also makes other research questions emerge. The different ways in which researchers can follow up on this work are:

1. **Dataset Expansion:** To increase the dataset and have a more diverse population for the study which in turn may lead to better generalizability of the findings.

2. **Feature Enhancement:** Adding data such as clinical markers and genetic data and lifestyle factors.

3. **Longitudinal Analysis:** Creating time-series models that predict the development of diabetes.

4. **Multi-Modal Learning:** To merge clinical data, with imaging information, and genomic data.

5. **Clinical Validation:** Organize future studies in which the clinical utility would be verified.

6. Mobile Integration: Design mobile applications that help with the real-time monitoring of risk factors.

7. Federated Learning: Carry out privacy-preserving training on a multi-healthcare institution basis.

6.3. Final Remarks

Demonstrating an accurate prediction with explainable AI working together is an essential milestone in the advent of reliable AI in the health sector. Although the technical aspect is critical, the real success of these systems will be their capability to increase the quality of the clinical decision-making process while being transparent and the ethical standards being respected.

Through technical innovations on the use of AI in the health sector, the focus on predictive performance and interpretability will always be vital. The framework we have defined shows that the two can be balanced and it is the launch pad for further advancements in decipherable medical AI.

ACKNOWLEDGEMENT

The authors would like to thank the UCI Machine Learning Repository for providing the PIMA Indians Diabetes dataset. We also acknowledge the open-source community for developing the machine learning and explainable AI libraries used in this research.

REFERENCES

- [1] International Diabetes Federation, "IDF Diabetes Atlas, 10th Edition," 2021. [Online]. Available: <https://www.diabetesatlas.org>
- [2] American Diabetes Association, "Economic Costs of Diabetes in the U.S. in 2021," *Diabetes Care*, vol. 45, no. 2, pp. 301-312, 2022.
- [3] World Health Organization, "Global Report on Diabetes," 2016. [Online]. Available: <https://www.who.int/diabetes/global-report>
- [4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8-17, 2015.
- [5] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Symposium on Computer Applications in Medical Care*, 1988, pp. 261-265.
- [6] M. L. Johnson, B. J. Smith, and A. K. Patel, "Ensemble methods for diabetes prediction using clinical and genetic data," *Journal of Biomedical Informatics*, vol. 89, pp. 1-12, 2019.
- [7] R. R. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
- [8] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765-4774.
- [9] J. W. Smith et al., "PIMA Indians Diabetes Database," UCI Machine Learning Repository, 1990. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>
- [10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.
- [11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.