Scientific Journal of Informatics
Vol. 12, No. 2, May 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Integration of Random Forest, ADASYN, and SHAP for Diabetes Prediction and Interpretation

Hozana Aulia[1*], Adi Wibowo[2], Sutrisno[3]
[1]Master of Information Systems, Postgraduate School, Universitas Diponegoro, Indonesia
[2]Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
[3]Department of Mathematics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia

Abstract.
Purpose: Diabetes is a chronic disease with a globally rising prevalence. Early detection of individuals at risk is essential to prevent long-term complications. This study aims to develop a diabetes prediction model that not only achieves high classification accuracy but also provides transparent explanations of the factors influencing its predictions.
Methods: The study utilized the Pima Indians Diabetes Dataset, which contains clinical data from 768 female patients aged over 21. The methodology included data preprocessing (handling of missing values and feature engineering, such as the creation of Age_BMI and Glucose_BMI features), a 70:30 train-test split, class imbalance handling using the ADASYN technique, model development using the Random Forest algorithm with hyperparameter tuning via GridSearchCV, and model interpretability analysis using SHAP.
Result: The proposed model achieved an accuracy of 79.2% and a recall of 85.2% on the test data. SHAP analysis revealed that Glucose, Age_BMI, BMI, and DiabetesPedigreeFunction were the most influential features in predicting diabetes. Furthermore, the SHAP heatmap indicated that individuals aged 30–50 years with obesity were at the highest risk. These findings align with existing medical literature, reinforcing the role of metabolic and age-related factors in diabetes development.
Novelty: This study presents an integrative approach combining class balancing (ADASYN), classification (Random Forest), and model interpretability (SHAP) in a unified framework for diabetes prediction. It emphasizes the importance of transparent model interpretation for healthcare professionals, enabling not only predictive outcomes but also actionable insights into risk factors. The findings support future research opportunities, including the integration of lifestyle variables and external validation using real-world clinical data from diverse populations.

Keywords: SHAP, ADASYN, Random forest, Diabetes prediction, Machine learning
Received May 2025 / **Revised** May 2025 / **Accepted** June 2025

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder with a steadily increasing global prevalence. According to the World Health Organization (WHO), over 537 million adults were diagnosed with diabetes in 2021, and this number is projected to rise significantly in the coming decades [1]. Beyond its high prevalence, diabetes is associated with a range of serious complications, including cardiovascular disease, kidney failure, and vision impairment, all of which significantly reduce patients' quality of life and impose a substantial burden on healthcare systems. Therefore, early detection and accurate prediction of diabetes are essential for supporting timely and appropriate medical decision-making [2].

In recent years, Machine Learning (ML) has emerged as a promising approach for automating disease detection through data-driven analysis [3]. However, the development of predictive ML models in healthcare continues to face several challenges. First, medical datasets—such as the widely used Pima Indians Diabetes Dataset typically exhibit class imbalance, where the number of positive cases (diabetic patients) is significantly lower than negative cases [4] Second, the presence of incomplete or noisy data, due to variability in clinical measurements, complicates model training and generalization [5]. Third, many high-performing ML models function as "black boxes," providing limited insight into the reasoning behind their predictions, which hinders their practical adoption in clinical settings [6].

To address these limitations, this study proposes an integrated ML-based diabetes prediction framework that combines three core components: Random Forest (RF), Adaptive Synthetic Sampling (ADASYN), and SHapley Additive exPlanations (SHAP). Random Forest was selected for its ability to handle

*Corresponding author.
Email addresses: hozanaaulia2471@gmail.com (Aulia), adiwibowo@lecturer.undip.ac.id (Wibowo), s.sutrisno@live.undip.ac.id (Sutrisno)
DOI: 10.15294/sji.v12i2.24314

heterogeneous clinical data, resistance to overfitting, capacity to manage noisy or missing values, and capability to capture non-linear interactions between features [7] Random Forest has also demonstrated superior classification performance compared to several other algorithms in various disease prediction studies[8].

ADASYN was used to address the class imbalance by generating synthetic samples for the minority class adaptively, especially in regions where samples are more difficult to classify [9]. Compared to conventional oversampling techniques such as SMOTE, ADASYN has shown better performance in improving model sensitivity [10] SHAP was employed to overcome the challenge of model interpretability by assigning a quantitative contribution score to each feature based on the Shapley value principle from cooperative game theory. His method allows both local and global explanations of model predictions, enabling medical professionals to validate and trust the outcomes [11].

Although numerous studies have applied ML to diabetes prediction, most have focused primarily on improving accuracy, while neglecting two critical aspects: class imbalance and model interpretability. For instance, earlier studies using Support Vector Machines, Logistic Regression, or deep learning models often lacked interpretability mechanisms, making their predictions difficult to verify in clinical practice [12] [13] Other studies have addressed imbalance using oversampling techniques like SMOTE, but without integrating interpretability tools, leaving uncertainty in model transparency [14].

This study addresses these research gaps by proposing a comprehensive and integrated approach that emphasizes both predictive performance and interpretability. By combining Random Forest as a robust classifier, ADASYN for adaptive class balancing, and SHAP for transparent feature attribution, the proposed model aims to deliver accurate and trustworthy predictive outcomes. The use of the Pima Indians Diabetes Dataset, a widely accepted benchmark in academic research, also enables comparison with prior work and highlights the novelty of the proposed approach. To highlight the contribution and novelty of this study, Table 1 compares previous works on diabetes prediction in terms of features and methods used. Most prior studies focused on improving accuracy but did not integrate class balancing and interpretability. Addressing this gap, this study proposes an integrated approach combining Random Forest for classification, ADASYN for adaptive class balancing, and SHAP for feature-level interpretability. The use of the widely accepted PIMA Indians Diabetes Dataset supports comprehensive evaluation and facilitates comparison with existing research.

Table 1. Comparison of features and methods used in previous diabetes prediction studies

Reference	Glucose	BMI	Age	Feature Engineering	Method
Febrian et al. (2023)	✓	✓	✓	×	Naïve Bayes, KNN
Kumari et al. (2021)	✓	✓	✓	×	Soft Voting (Random Forest, Naïve Bayes, Logistic Regression)
Chatrati et al. (2022)	✓	×	×	×	Support Vector Machine
This study	✓	✓	✓	✓	Random Forest + ADASYN + SHAP

As observed in Table 1, previous works such as those by Febrian et al. (2023) and Kumari et al. (2021) utilized glucose, BMI, and age as input features but did not apply feature engineering. Their models relied on conventional classifiers, including Naïve Bayes, K-Nearest Neighbors, and ensemble voting techniques. Chatrati et al. (2022) adopted a more limited approach by using only glucose and applying a Support Vector Machine (SVM), also without any feature transformation. These studies primarily focused on improving classification performance, with limited consideration for class imbalance or model interpretability. In contrast, the present study integrates Random Forest for classification, ADASYN for adaptive class balancing, and SHAP for feature-level interpretability, offering a more comprehensive and explainable approach to diabetes prediction.

METHODS

This study uses a machine learning approach to build a diabetes prediction model. Each stage of the method is designed to ensure optimal model performance while maintaining transparency in the interpretation of predicted results [15]. The proposed methodological workflow is illustrated in Figure 1. The process begins with loading the PIMA Indians dataset, followed by preprocessing steps such as data cleaning and feature engineering. The dataset is then divided into training and testing subsets. To address the class imbalance, ADASYN oversampling is applied to the training data only [16]. A Random Forest model is trained and

used to predict patient outcomes as diabetic or non-diabetic. Model performance is evaluated using accuracy, precision, recall, and F1-score. To improve interpretability, SHAP is applied to explain the contribution of each feature to the predictions [17].

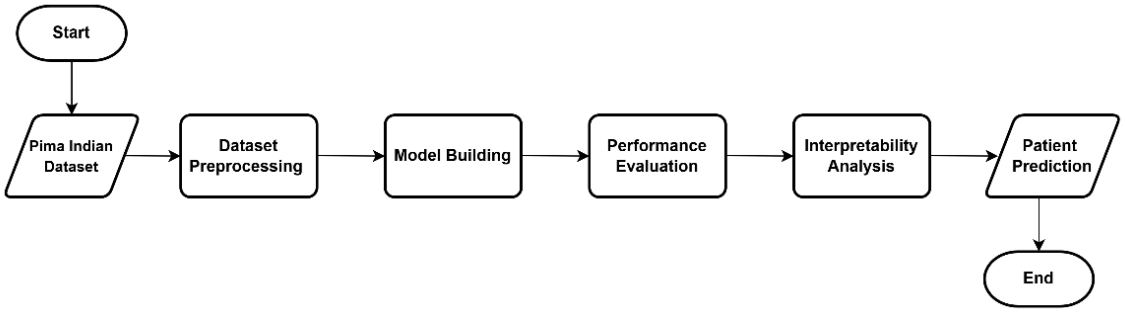

Figure 1. Proposed method flowchart

Diabetes Dataset

This research utilizes PIMA Indians data which can be accessed freely through the Kaggle platform (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). Table 2 shows the composition of the attributes used in the diabetes dataset. This dataset consists of 768 data on patients who are female, with an age range between 21 and 81 years. Of the total data, 500 data came from individuals without diabetes, while the other 268 data represented individuals who were diagnosed with diabetes. This dataset has eight (8) main features, including the number of pregnancies, glucose levels, diastolic blood pressure, thickness of triceps skin folds, insulin levels after two hours, body mass index (BMI), history of hereditary diseases, and age Despite its popularity as a benchmark dataset, it has demographic limitations, as all subjects are female from a single ethnic group (Pima Indian). This may introduce bias and limit the generalizability of the model. Further validation on more diverse clinical datasets is recommended to ensure broader applicability.

Table 2. Description of the Pima Indians diabetes dataset

Feature Name	Data Type	Value of Data	Number of Unique
Pregnancies	Integer	0-17	18
Glucose	Integer	0-199	136
BloodPressure	Integer	0-122	47
SkinThickness	Integer	0-99	51
Insulin	Integer	0-846	186
BMI	Float	0.0 – 67.1	248
DiabetesPedigreeFunction	Float	0.078 – 2.42	517
Age	Integer	21 – 81	52
Outcome (Label)	Integer	0 (No), 1 (Yes)	2

Data Distribution

Figure 2 shows the histograms of all numerical features in the dataset. Some features, such as Insulin, Age, and DiabetesPedigreeFunction, are heavily skewed to the right, while others like Glucose and BMI show mild skewness. These uneven distributions, along with the imbalance between diabetic and non-diabetic cases in the target variable, support the use of ADASYN. This method helps balance the data and improves learning in areas where the data is sparse and harder for the model to classify.

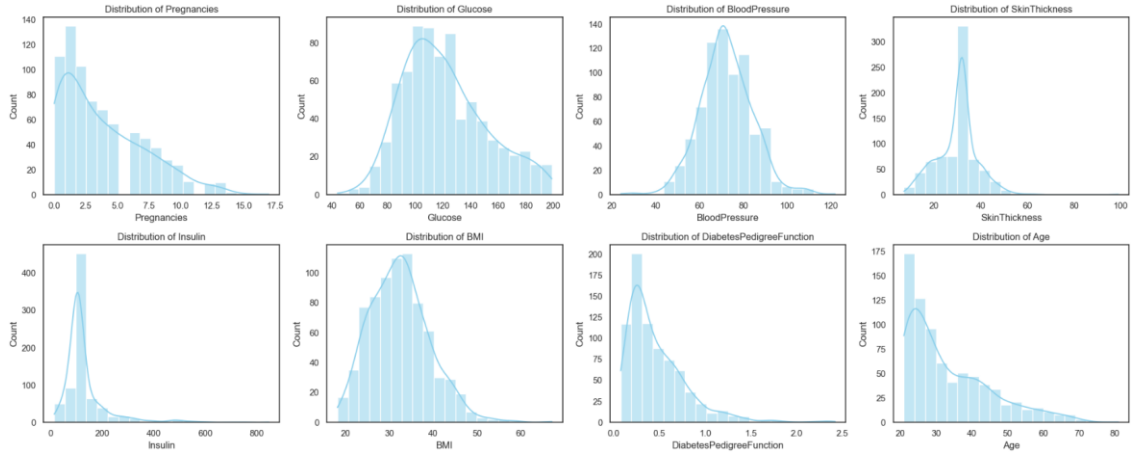

Figure 2. Data distribution

Preprocessing Data

This stage aims to ensure that the quality of the data used is good and ready to build the model [18]. This data preprocessing involves steps ranging from data cleaning, feature engineering, class distribution checks, and correlations between variables.

Missing Value Handling

At this stage, medical features that are logically unlikely to be worth 0, such as glucose, diastolic blood pressure, skin thickness, insulin, and BMI are identified. The value of 0 is categorized as hidden missing values and replaced with NaN. Imputation is then carried out using the median of each feature to maintain the consistency of data distribution and reduce the influence of outliers [19].

Feature Engineering

To enrich the dataset, two new features were added, namely, glucose_BMI (ratio of glucose to body mass index) and Age_BMI (multiplication of age by body mass index). These engineered features aim to capture potential nonlinear interactions between metabolic and demographic variables that may contribute to diabetes risk. The complete list of features used in the modeling process, along with the additions resulting from the feature engineering step, is presented in Table 3.

Table 3. Comparison of features before and after feature engineering

Original Features	Feature Engineering
Pregnancies	Pregnancies
Glucose	Glucose, **Glucose_BMI (New)**
BloodPressure	BloodPressure
SkinThickness	SkinThickness
Insulin	Insulin
BMI	BMI, **AGE_BMI (New)**
DiabetesPedigreeFuction	DiabetesPedigreeFuction
Age	Age

Correlation Between Variables

Based on the correlation heatmap in Figure 3, Glucose showed the highest correlation to Outcome (0.49), confirming its role as a key feature of prediction. In addition, Glucose_BMI and Age_BMI engineering features have a strong correlation with their constituent variables, supporting the validity of the feature engineering carried out.

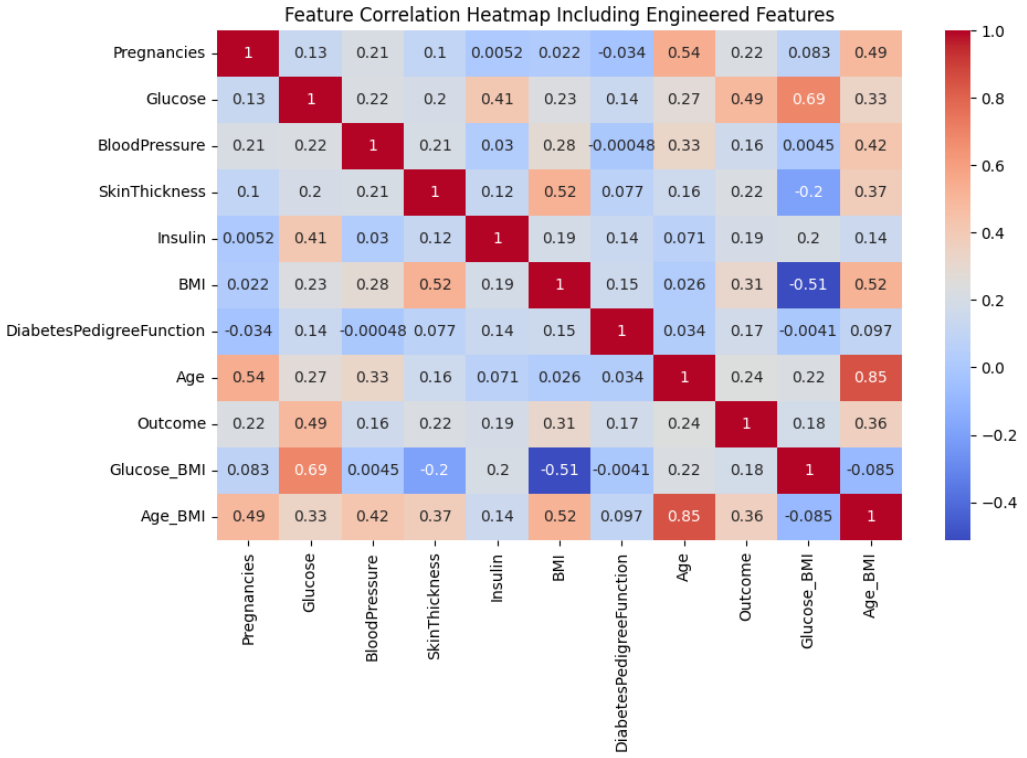

Figure 3. Correlation between variables

Split Data

After the preprocessing stage, the dataset is separated into features (X) and targets (y). Feature (X) consists of Glucose, BMI, Insulin, Age, Pregnancies, BloodPressure, Glucose_BMI, Age_BMI, Skinthickness, and DiabetesPedigreeFunction, while target (y) is Outcome. The data was then divided into 70% for training (537 data) and 30% for testing (231 data).

Balancing Data

Based on the class distribution, the dataset showed an imbalance, with the number of individuals without diabetes much greater than those with diabetes. This imbalance has the potential to lower the model's performance due to its bias against the majority class. To overcome this, this study applied the ADASYN oversampling technique [20] to improve accuracy and reduce classification bias. Figure 4 comparison of class label distributions before and after applying ADASYN. The minority class (label 1) was synthetically increased to match the majority class, improving dataset balance.

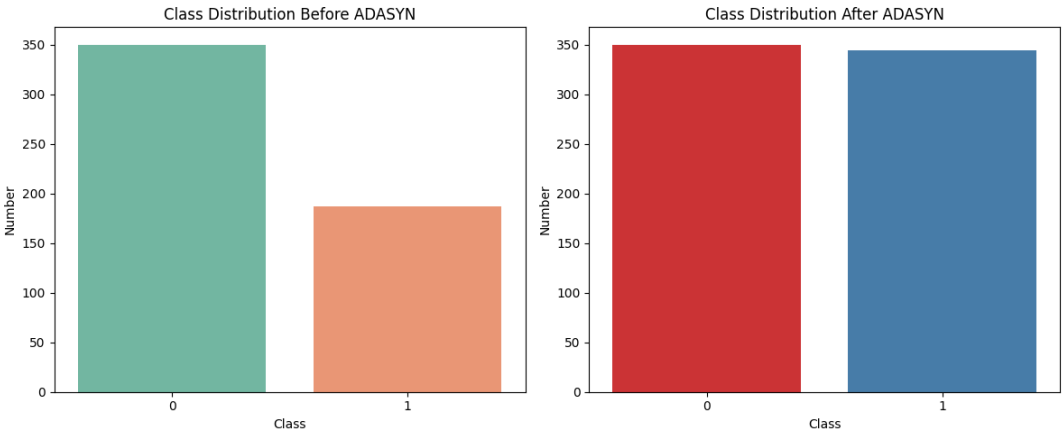

Figure 4. Class distribution before and after applying ADASYN

Model Development
At this stage, the model is built to study patterns from the trained data that has gone through pre-processing and feature engineering. The model developed aims to be able to generalize these patterns in order to make accurate predictions of new data that have never been seen before [21].

The development of the model in this study was carried out with the following approach. The algorithm used is Random Forest, which is a decision tree-based ensemble method that is known to have high classification performance and resistance to overfitting [22]. To overcome the imbalance of class distribution in the dataset, the ADASYN oversampling technique was applied, which aims to increase the representation of minority classes in the training data so that the model can learn in a more balanced manner [23].

The optimization process was performed using GridSearchCV[15], [24] combined with Stratified K-Fold Cross-Validation (K=5) to identify the best combination of hyperparameters. The tuned parameters included n_estimators (number of trees), max_depth (maximum depth of each tree), min_samples_split (minimum number of samples required to split an internal node), and min_samples_leaf (minimum number of samples required at a leaf node). For max_features, the options 'sqrt' and 'log2' were tested, corresponding to the square root and logarithm (base 2) of the total number of features, respectively. These settings help control model complexity and reduce overfitting. The selection of the best configuration was based on accuracy as the scoring metric. To improve computational efficiency, the search was executed in parallel using n_jobs=-1, and progress was monitored with verbose=2. Once the training process was completed, the best model obtained from GridSearchCV was evaluated on the test dataset.

Model Evaluation
Once the model has been built and trained on the training data, the next stage involves evaluating its performance in accurately classifying unseen data. The main objective of this evaluation is to assess the model's generalization ability using the test dataset. Several performance metrics are employed, including accuracy, precision, recall, and F1-score, which are particularly important in reflecting balanced performance on imbalanced datasets. In addition, a confusion matrix is used to provide a detailed breakdown of true and false predictions for each class.

Interpretability with SHAP
To improve the interpretability of the model, this study uses SHAP. SHAP serves to explain the contribution of each feature to model prediction fairly, based on the Shapley Value theory of game theory [5]. In this study, TreeExplainer was used to calculate the SHAP value on the trained Random Forest classifier. TreeExplainer leverages the internal structure of decision trees to compute exact Shapley values efficiently without resorting to costly sampling or permutation methods [25]. This makes it suitable for large datasets and ensemble models like Random Forest.

SHAP values were computed for each instance in the test set to provide both global interpretability identifying overall feature importance across the dataset and local interpretability, which explains how individual features contribute to specific predictions. Visualizations including summary plots, dependence plots, and force plots were used to illustrate model behavior and highlight the features that most strongly influenced diabetes classification outcomes.

This interpretability layer is particularly important in medical contexts, as it enables healthcare professionals to not only observe the model's predictions but also understand the underlying rationale. Such transparency is essential for building trust and supporting the integration of AI-assisted tools in clinical decision-making.

RESULTS AND DISCUSSIONS

Model Development Results
The diabetes classification model was developed using the Random Forest algorithm, combined with the ADASYN oversampling technique to address class imbalance in the dataset. Hyperparameter tuning was conducted using GridSearchCV in conjunction with 5-fold Stratified Cross-Validation. The optimal hyperparameters obtained from the tuning process are summarized in Table 4, which include: 250 trees (n_estimators), a maximum tree depth of 20 (max_depth), a minimum of 10 samples required to split an internal node (min_samples_split), and a minimum of 20 samples at a leaf node (min_samples_leaf).

Additionally, the number of features considered at each split (max_features) was set to either 'sqrt' or 'log2', two commonly used strategies that help improve model efficiency while minimizing the risk of overfitting.

Table 4. Hyperparameters with GridSearchCV

Parameter	Value
n_estimators	250
max_depth	20
min_samples_split	10
min_samples_leaf	20
max_features	*'sqrt', 'log2'*

Model Performance Analysis

The classification performance of the model, detailed in Table 5, shows an accuracy of 79.2%, a recall of 85.2%, a precision of 66%, and an F1-score of 74.2%. The high recall reflects the model's strong ability to identify diabetic cases, largely attributed to the use of ADASYN, which improved minority class representation during training, and the Random Forest classifier's sensitivity to key clinical features such as glucose and BMI.

However, the lower precision indicates a trade-off, with more false positives produced as the model prioritized sensitivity. In clinical settings, recall is often prioritized, especially in screening scenarios, as false negatives (undiagnosed diabetic patients) carry greater risk than false positives, which can be resolved through follow-up testing. This trade-off aligns with prior findings in medical classification tasks, where high recall is considered more valuable for screening conditions with serious health consequences.

Thus, the strategy of emphasizing recall through data balancing and model selection is appropriate for early detection purposes, where minimizing missed cases is more critical than optimizing overall accuracy.

Table 5. Model classification results

Aspect of Evaluation	Value	Information
Accuracy	79.2%	The percentage of correct predictions for all test data
Precision	66%	The model's ability to accurately predict diabetes cases
Recall	85.2%	The model's ability to detect all existing diabetes cases
F1-Score	74.2%	The balanced average of precision and recall for diabetes classification

Confusion Matrix Analysis

Figure 5 presents the confusion matrix generated from the model's predictions on the test set, offering a detailed breakdown of classification outcomes. The model correctly identified 114 diabetic individuals, referred to as true positives (TP). However, it also misclassified 36 diabetic patients as non-diabetic, resulting in false negatives (FN) critical concern in medical settings, as these patients may not receive the necessary treatment or monitoring. On the other hand, the model accurately recognized 69 non-diabetic individuals, recorded as true negatives (TN). Additionally, 12 non-diabetic cases were incorrectly predicted as diabetic, and categorized as false positives (FP). Although false positives may lead to unnecessary follow-up examinations, they are generally less harmful than false negatives, which carry a higher risk in clinical decision-making.

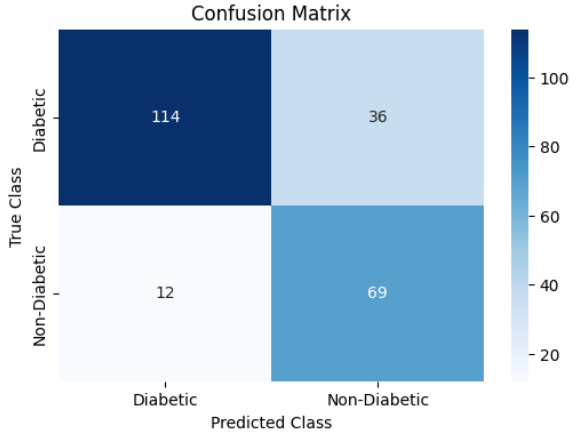

Figure 5. Confusion matrix with true class and predicted class

Cross Validation and Generalization

To assess generalization, a 5-fold stratified cross-validation was conducted using the best hyperparameters from GridSearchCV. The model obtained a mean accuracy of 76.6% ± 3.8%, indicating consistent performance across different data splits. On the test set, the model achieved an accuracy of 79.2%, with a precision of 66%, a recall of 85.2%, and an F1-score of 74.2%. The small gap between cross-validation and test performance suggests that the model generalizes well without overfitting. The high recall is particularly valuable in clinical settings to minimize missed diabetes cases.

Comparative with Previous Studies

Table 6 summarizes the performance of various machine learning models from previous studies on the Pima Indians Diabetes dataset, and compares them to the model developed in this study.

Table 6. Comparative evaluation of machine learning models

Reference	Method	Accuracy	Precision	Recall	F1-Score
Chatrati et al. (2023)[5]	SVM	75%	-	-	-
Kumari et al. (2021)[26]	SoftVoting (RF+NB+LR)	79.08%	73.1%	70%	71.5%
Febrian et al. (2024)[2]	KNN, NB	76%	73%	71%	-
This study	ADASYN+RF+SHAP	79.2%	66%	85.2%	74.1%

Table 6 demonstrates that the proposed model (ADASYN+RF+SHAP) achieves competitive performance, with the highest accuracy (79.2%) and recall (85.2%) among the compared studies. Although the precision is slightly lower, the balanced F1-score (74.1%) indicates that the model is effective, particularly in handling imbalanced data and improving the detection of positive diabetes cases.

Model Interpretation with SHAP

SHAP Summary Plot

To understand the factors influencing the model's predictions, an interpretability analysis was performed using SHAP. This method quantifies the contribution of each feature to the prediction outcome, enabling a transparent and interpretable explanation of the model's decisions.

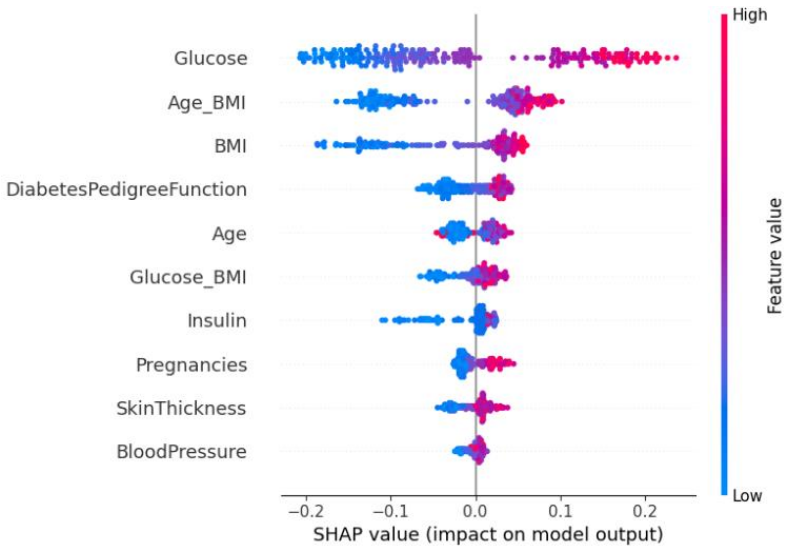

Figure 6. SHAP Summary Plot

Figure 6 presents a SHAP summary plot that illustrates the contribution of each feature to the diabetes prediction outcome. The Y-axis lists the feature names, while the X-axis displays the SHAP values, which indicate both the magnitude and direction of each feature's contribution. The color of each dot represents the actual value of the feature, ranging from low (blue) to high (red).

Among all features, Glucose contributes the most significantly, with higher values associated with an increased risk of diabetes (positive SHAP values). Age_BMI and BMI also exhibit strong influence,

followed by DiabetesPedigreeFunction, Age, and Glucose_BMI. In contrast, features such as Insulin, Pregnancies, SkinThickness, and BloodPressure contribute less to the model's predictions. This global interpretability analysis highlights the dominant predictors and supports clinical understanding of the model's decision-making process.

Local SHAP Explanation of Individual Prediction

Figure 7 displays a SHAP waterfall plot that explains the prediction for a single patient. The model's baseline output was 0.497. Two features contributed to the final prediction: Glucose (+0.03) increased the risk, indicating that the patient's glucose level was above average, while Pregnancies (–0.03) reduced the risk, suggesting a lower number of pregnancies compared to typical diabetic cases. The final prediction remained at 0.497, which is close to the classification threshold, indicating a borderline case. This local explanation enhances model transparency and helps clinicians understand the key factors influencing the model's decision at the individual level.

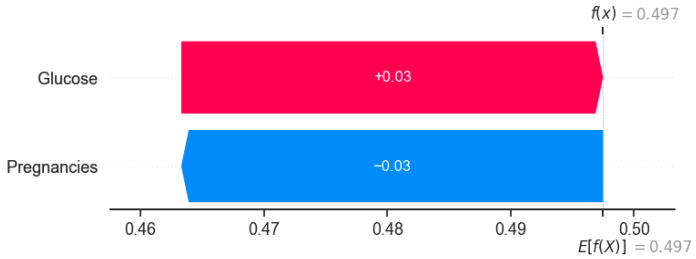

Figure 7. Local SHAP Explanation

Analysis of Average SHAP Scores by Age Group and BMI Category

To gain a deeper understanding of the interaction between age and body mass index (BMI) in model predictions, an analysis was conducted on the average SHAP values across different age groups and BMI categories. This analysis aims to reveal patterns in feature contributions across population subgroups, providing a more contextual and clinically meaningful interpretation of the model's predictions in diabetes classification.

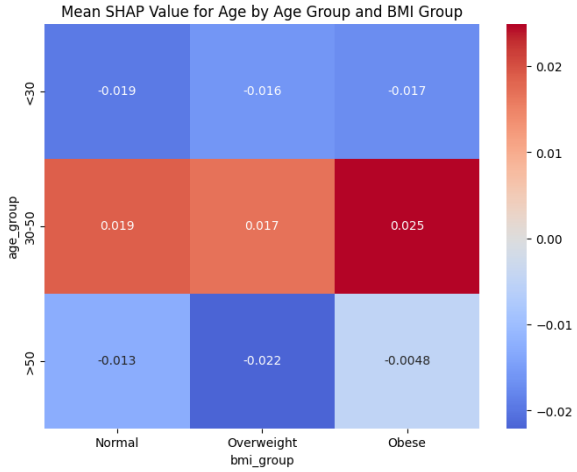

Figure 8. Subgroups by Age

Figure 8 presents a heatmap of the average SHAP values for the Age feature across age groups (<30, 30–50, and >50 years) and BMI categories (Normal, Overweight, Obese). The results indicate that the 30–50 age group with obesity contributes the most to diabetes prediction (average SHAP value: 0.025), suggesting that age plays a significant role in increasing the likelihood of a positive classification within this subgroup. In contrast, the contribution of age in individuals under 30 or over 50 is relatively low, and in some cases negative, indicating a lesser impact of this feature in those segments.

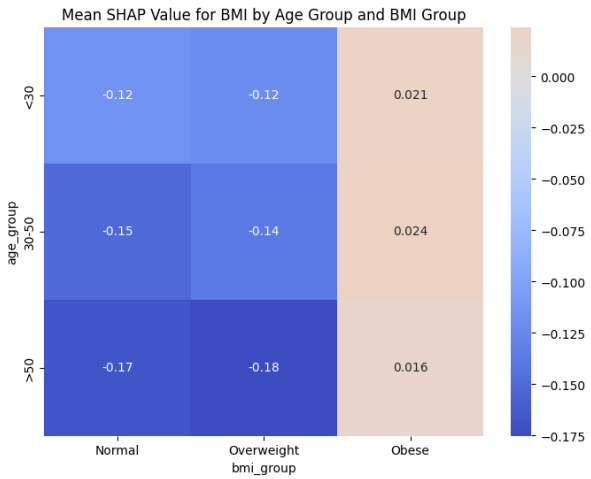

Figure 9. Subgroups by BMI

Figure 9 displays a heatmap of the average SHAP values for the BMI feature using the same dimensions. The highest values were again observed in the 30–50-year age group with obesity (SHAP value: 0.024), indicating that BMI significantly increased the model's prediction of diabetes in this subgroup. In contrast, notably negative SHAP values were observed among individuals with normal or overweight BMI, particularly those aged over 50, suggesting that BMI in non-obese categories tends to reduce the model's likelihood of predicting diabetes.

Overall, this analysis demonstrates that feature contributions to model predictions are not uniform but are influenced by the interaction between age and weight status. These findings not only enhance the medical relevance of the model but also highlight the importance of data stratification in interpreting machine learning-based predictions.

CONCLUSION

This study developed an integrated ADASYN–Random Forest–SHAP model that achieved 79.2% accuracy and 85.2% recall on the PIMA Indians dataset. SHAP analysis highlighted Glucose, Age_BMI, and BMI as key predictors, with the highest risk observed among obese individuals aged 30–50 years. While the model shows effective classification performance and interpretability, its generalizability is limited by the dataset's demographic scope and the absence of external validation. Future work should incorporate lifestyle variables and test the model on more diverse clinical data to support real-world deployment in decision support systems.

REFERENCES

[1] International Diabetes Federation (IDF), "IDF Diabetes Atlas," Brussels, Dec. 2025.

[2] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 21–30. doi: 10.1016/j.procs.2022.12.107.

[3] G. S, R. Venkata Siva Reddy, and M. R. Ahmed, "Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus," *Measurement: Sensors*, vol. 31, Feb. 2024, doi: 10.1016/j.measen.2023.100983.

[4] A. B. Devi, "The Effect of Anomaly Detection and Data Balancing in Prediction of Diabetes," 2024.

[5] S. P. Chatrati *et al.*, "Smart home health monitoring system for predicting type 2 diabetes and hypertension," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 862–870, Mar. 2022, doi: 10.1016/j.jksuci.2020.01.010.

[6] Q. Fu *et al.*, "Identifying cardiovascular disease risk in the U.S. population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology," *Ecotoxicol Environ Saf*, vol. 286, Nov. 2024, doi: 10.1016/j.ecoenv.2024.117210.

[7] A. Khairunnisa, K. A. Notodiputro, and B. Sartono, "A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 207–218, Feb. 2024, doi: 10.15294/sji.v11i1.48721.

[8] L. Jiang *et al.*, "Diabetes risk prediction model based on community follow-up data using machine learning," *Prev Med Rep*, vol. 35, Oct. 2023, doi: 10.1016/j.pmedr.2023.102358.

[9] R. Mitra, A. Bajpai, and K. Biswas, "ADASYN-assisted machine learning for phase prediction of high entropy carbides," *Comput Mater Sci*, vol. 223, Apr. 2023, doi: 10.1016/j.commatsci.2023.112142.

[10] T. Xu, G. Coco, and M. Neale, "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning," *Water Res*, vol. 177, Jun. 2020, doi: 10.1016/j.watres.2020.115788.

[11] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-44155-x.

[12] B. M. Manjula, A. Harshavardhan, B. T. Indrasena, and S. Neha Annie, "Exploratory Data Analysis and Predictive Modeling of Pima Indian Diabetes," in *International Conference on Intelligent Algorithms for Computational Intelligence Systems, IACIS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IACIS61494.2024.10721874.

[13] L. Chandra Sekhar Reddy, M. Gottipalli, P. Sravanthi, J. Rajanikanth, G. Yalamarthi, and N. Gurrapu, "Bridging Horizons in Diabetes Prediction: A Comparative Exploration of Machine Learning and Deep Learning Approaches in Pima Indian Women," in *Proceedings - 2nd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 386–391. doi: 10.1109/InCACCT61598.2024.10550977.

[14] M. K. Rezki, M. I. Mazdadi, F. Indriani, Muliadi, T. H. Saragih, and V. A. Athavale, "Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 4, pp. 343–354, Oct. 2024, doi: 10.35882/jeeemi.v6i4.434.

[15] A. S. Antonini *et al.*, "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," Sep. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.acags.2024.100178.

[16] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 276–282, Nov. 2021, doi: 10.15294/sji.v8i2.32484.

[17] M. Fu, Y. Liu, Z. Hou, and Z. Wang, "Interpretable prediction of acute ischemic stroke after hip fracture in patients 65 years and older based on machine learning and SHAP," *Arch Gerontol Geriatr*, vol. 129, Feb. 2025, doi: 10.1016/j.archger.2024.105641.

[18] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," *IEEE Access*, vol. 12, pp. 80698–80730, 2024, doi: 10.1109/ACCESS.2024.3406748.

[19] V. Jain, S. Shukla, and N. Khare, "Analysis of various data imputation techniques for diabetes classification on PIMA dataset," in *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/SCEECS61402.2024.10482050.

[20] T. M. Khan, S. Xu, Z. G. Khan, and M. U. Chishti, "Implementing multilabeling, ADASYN, and relieff techniques for classification of breast cancer diagnostic through machine learning: Efficient computer-aided diagnostic system," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/5577636.

[21] D. Dai *et al.*, "Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys," *Comput Mater Sci*, vol. 175, Apr. 2020, doi: 10.1016/j.commatsci.2020.109618.

[22] E. Asamoah, G. B. M. Heuvelink, I. Chairi, P. S. Bindraban, and V. Logah, "Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana," *Heliyon*, vol. 10, no. 17, Sep. 2024, doi: 10.1016/j.heliyon.2024.e37065.

[23] E. Ileberi and Y. Sun, "Advancing Model Performance With ADASYN and Recurrent Feature Elimination and Cross-Validation in Machine Learning-Assisted Credit Card Fraud Detection: A Comparative Analysis," *IEEE Access*, vol. 12, pp. 133315–133327, 2024, doi: 10.1109/ACCESS.2024.3457922.

[24] G. S. Manivannan, H. Rajaguru, R. S, and S. V. Talawar, "Cardiovascular disease detection from cardiac arrhythmia ECG signals using artificial intelligence models with hyperparameters tuning methodologies," *Heliyon*, vol. 10, no. 17, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36751.

[25] P. Sharma *et al.*, "Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.06734

[26] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.