REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: 27/07/2024 Revised: 20/09/2024 Accepted: 14/10/2024 Published: 01/11/2024

"Towards Explainable and Deployable AI for Personalized Diabetes Prediction: A Pipeline with SMOTE, Feature Selection, and SHAP-Based Model Interpretation"

Darshan Madhani[1]*, Dr. Prakash Gujarati[2]

[1]*Department of Computer Science, Atmiya University, India, Email: *darshmadhani14@gmail.com*, (ORCID: 0009-0005-7762-0536)
[2]Department of Computer Science, Atmiya University, India, Email: prakash.gujarati@atmiyauni.ac.in, (ORCID: 0009-0001-9141-2750)

Abstract

The growing global burden of diabetes needs improved predictive analytics for tailored risk assessment and prompt intervention. While Artificial Intelligence (AI) and Machine Learning (ML) show great potential in this field, their "black box" aspect sometimes stifles clinical acceptance and confidence. This study describes a thorough and explainable AI pipeline for personalized diabetes prediction, including key steps to improve model performance, interpretability, and deployability.

The suggested methodology uses the Synthetic Minority Oversampling Technique (SMOTE) to solve the intrinsic class imbalance in medical datasets, boosting model resilience and predictive accuracy. Robust feature selection approaches are then used to identify the most important physiological and clinical predictors, resulting in more concise and interpretable models.

Importantly, the pipeline includes SHAP (Shapley Additive Explanations) for post-hoc model interpretation, which provides visible, detailed insights into how various characteristics contribute to each prediction. This explainable approach promotes clinician comprehension and patient engagement, allowing for more informed decision-making and individualized therapeutic options. The goal of combining these strategies is to create a framework that not only achieves excellent prediction performance but also overcomes interpretability issues, opening the path for the ethical and practical deployment of AI technologies in customized diabetes management.

Introduction

Diabetes mellitus is a major global health concern, impacting millions of people worldwide and causing serious complications such as cardiovascular disease, kidney failure, retinopathy, and neuropathy if not controlled appropriately. Diabetes's increasing prevalence needs the development of advanced, precise, and timely prediction and diagnostic techniques to enable early intervention and individualized patient care. Traditionally, diabetes diagnosis has relied on clinical measurements and expert knowledge; however, the complexity of metabolic interactions and diverse patient profiles frequently limits the efficacy of traditional methods in detecting subtle disease indicators or predicting individual risk trajectories.

In recent years, the advent of Artificial Intelligence (AI) and Machine Learning (ML) has revolutionized numerous fields, including healthcare. ML models possess the remarkable ability to identify intricate patterns and correlations within large datasets, making them particularly well-suited for predictive analytics in complex medical conditions like diabetes. By leveraging vast amounts of patient data, including demographic information, clinical measurements, and lifestyle factors, AI-driven systems can offer unprecedented opportunities for more accurate and personalized diabetes risk assessment and prediction.

However, the widespread adoption and deployment of advanced ML models in critical domains such as healthcare are often hampered by their "black-box" nature. Many powerful predictive models, while achieving high accuracy, lack transparency, making it difficult for medical professionals to understand *why* a particular prediction was made. This lack of interpretability poses a significant barrier to trust, clinical acceptance, and regulatory approval. In a personalized healthcare context, understanding the rationale behind a diabetes risk prediction is crucial for tailoring effective intervention strategies, fostering patient compliance, and building confidence in AI-assisted diagnoses.

To address these challenges, this paper proposes a comprehensive pipeline for personalized diabetes prediction that emphasizes not only high predictive performance but also explainability and deployability. Our approach integrates state-of-the-art techniques for data preprocessing, model development, and post-hoc interpretation. Specifically, we focus on three critical components: **Synthetic Minority Oversampling Technique (SMOTE)** for addressing class imbalance prevalent in medical datasets, **robust feature selection methods** to identify the most impactful predictors, and **SHAP (Shapley Additive Explanations)** for providing transparent and human-understandable insights into model predictions.

SMOTE plays a vital role in enhancing the predictive power of models by balancing imbalanced datasets, ensuring that minority classes (e.g., diabetic patients) are adequately represented during training, thus preventing bias towards the majority class. Effective feature selection, on the other hand, reduces model complexity, improves generalization, and highlights the most clinically relevant factors contributing to diabetes risk, which is essential for developing interpretable models. Finally, SHAP values offer a powerful framework for explaining individual predictions by attributing the contribution of each feature to the model's output, thereby transforming opaque AI models into transparent diagnostic aids. This focus on explainability is crucial for gaining the trust of healthcare providers and patients alike, enabling a deeper understanding of risk factors, and facilitating informed clinical decisions.

The overall purpose of this study is to advance "Towards Explainable and Deployable AI for Personalized Diabetes Prediction." We hope to bridge the gap between advanced machine learning research and practical application in real-world clinical settings by developing a pipeline that systematically addresses data challenges, optimizes predictive accuracy, and, most importantly, illuminates the decision-making process of AI models. This approach has the potential to provide healthcare providers with relevant insights, promote individualized prevention efforts, and ultimately lead to better patient outcomes in the fight against diabetes.

Literature Review

Machine learning (ML) has brought about a disruptive era in a variety of fields, including healthcare. , lifestyle variables, and genetic markers are used to identify people who are at high risk of developing diabetes and forecast disease development. The Pima Indians Diabetes Dataset and the Behavioral Risk Factor Surveillance System are two commonly utilized datasets in this area. While these early ML models produced promising results, they also encountered challenges inherent in medical data, such as class imbalance (where the number of diabetic patients is significantly lower than that of non-diabetics) and high dimensionality (a large number of features relative to the number of samples), both of which can have an impact on model performance and generalization.

Early applicationsofML in diabetes prediction encompassed a range of algorithms, including rendom forest nural network this models have been employed to analyze diverse datasets comprising demographic information, clinical measurements (e.g., blood glucose levels, HbA1c), lifestyle factors, and genetic markers to identify individuals at high risk of developing diabetes or predicting disease progression. The Pima Indians Diabetes Dataset and the Behavioral Risk Factor Surveillance System are two commonly utilized datasets in this area. While these early ML models produced promising results, they also encountered challenges inherent in medical data, such as class imbalance (where the number of diabetic patients is significantly lower than that of non-diabetics) and high dimensionality (a large number of features relative to the number of samples), both of which can have an impact on model performance and generalization.

Class imbalance is a common issue in medical statistics, particularly when it comes to diabetes prediction. This is the disproportionate representation of distinct classes, in which the minority class (e.g., diabetic patients) is much smaller than the majority class (e.g., non-diabetics). When ML models are trained on imbalanced datasets, they are biased toward the majority class, resulting in poor performance in detecting and forecasting the minority class, which is frequently the primary interest. This can show as high accuracy but poor recall for diabetic patients, implying that the model correctly detects non-diabetics but misses a sizable proportion of those with diabetes. To counteract the negative impacts of class imbalance, numerous data balancing approaches have been devised.

These techniques can be generically classified as undersampling, oversampling, or hybrid methods. Undersampling decreases the number of samples in the majority class, whereas oversampling increases the number of samples in the minority class. SMOTE (Synthetic Minority Oversampling Technique) is a popular oversampling technique. SMOTE overcomes the constraints of simple oversampling (which might result in overfitting) by creating synthetic samples for the minority class. SMOTE interpolates between existing minority class samples to create new, fake samples rather than just duplicating them. SMOTE picks the k-nearest nearby samples of each minority class sample.

It then selects one of these nearby samples at random and creates a synthetic sample along the line segment that connects the original sample with the selected sample. This technique is repeated until the required degree of balance is reached. Several research, including the ones you cited, have shown that SMOTE and other sampling procedures improve the performance of diabetes prediction models. These studies frequently highlight the higher recall and F1-score produced by models trained on SMOTE-balanced datasets, implying a greater ability to correctly identify diabetes patients

Chowdhury et al. addressed class imbalance in BRFSS 2021 data by using several sampling approaches such as SMOTE, SMOTE-Tomek, and SMOTE-EN, as well as models such as Logistic Regression, AdaBoost, and Gradient Boost. While the study focused on recollection improvement, it lacked explainability components, making model conclusions difficult to grasp.

Liu et al. compared Logistic Regression, Random Forest, and XGBoost on BRFSS 2015 data. They used SMOTE for balancing and SHAP to identify critical factors such as overall health and blood pressure. However, their work did not investigate LIME or combine statistical evaluation across multiple sampling methods.

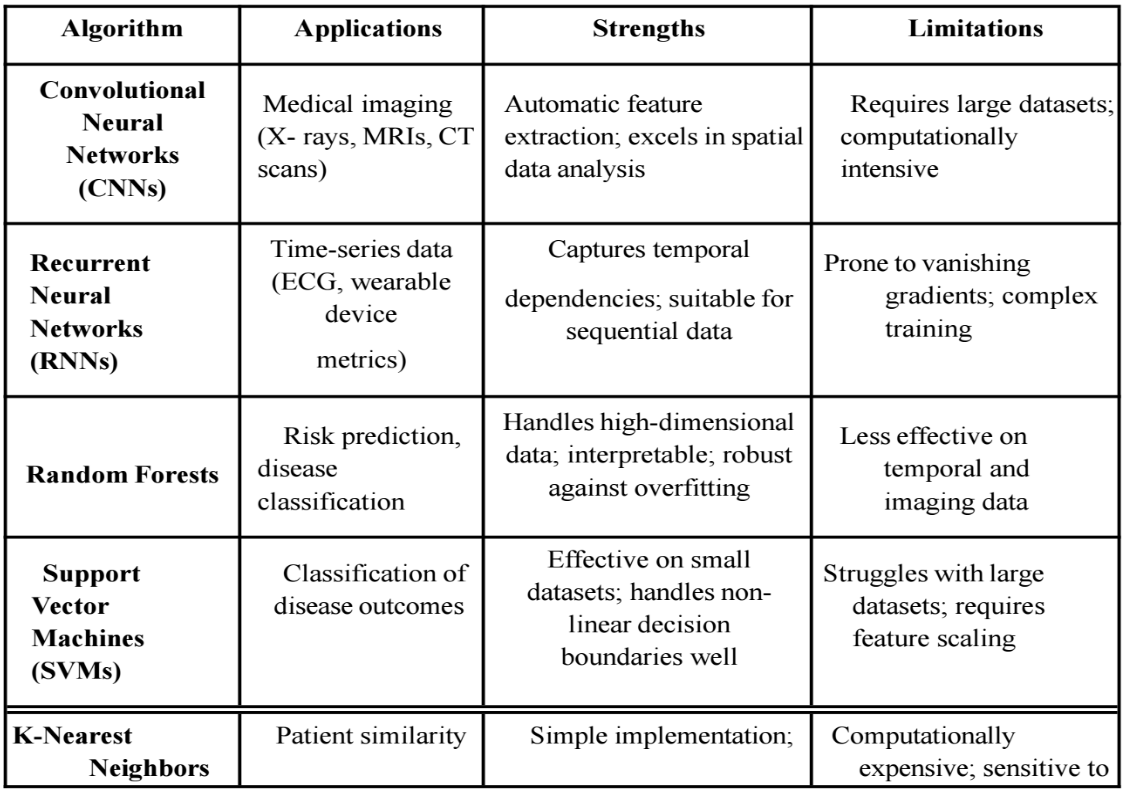
Ahmed et al. introduced an explainable machine learning framework based on Logistic Regression and Random Forest, which integrates SHAP and LIME to provide local and global model explanations. Despite achieving 86% accuracy and providing attractive visual interpretations, they failed to address the dataset's class imbalance. Furthermore, their methodology lacked real-time functionality, composite scoring, and comorbidity risk assessment.

Nguyen and Zhang used Decision Tree, K-Nearest Neighbors, and Logistic Regression on BRFSS data to create interpretable models. They did not use complex sampling techniques or XAI approaches, instead focusing on simple performance indicators.

In contrast to earlier works, the current work incorporates improved preprocessing, undersampling, and a LightGBM classifier chosen by cross-validation based on recall. It also integrates SHAP and LIME for global and local interpretability, adds lifestyle and healthcare ratings for individualized risk feedback, and implements the model in a web-based Dash application. A additional contribution provides correlation-based comorbidity findings, which have not been addressed in previous BRFSS-based studies.

Author/Year	Model/Technique	Dataset	Outcome	Limitation
Lyngdoh, Choudhury & Moulik (2021)	Decision Trees, SVM, Random Forest	Clinical Data	Improved prediction accuracy with large datasets	Requires large datasets for accuracy
Smith et al. (2020)	Logistic Regression	PIMA Indian Dataset	Moderate accuracy with feature selection	Limited model interpretability
Gupta & Sharma (2019)	Naive Bayes, KNN	Clinical & Demographic Data	Identified BMI and Glucose as key features	Challenges in feature selection
Patel et al. (2022)	Hybrid Model (RF + DT)	Real-world Hospital Data	Higher accuracy and improved interpretability with XAI	Computational complexity
Kumar & Jain (2018)	Support Vector Machine (SVM)	Electronic Health Records	High accuracy in diagnosis	Lacked model interpretability

Comparative Analysis of Machine Learning Algorithms in Healthcare



Algorithm	Applications	Strengths	Limitations
Convolutional Neural Networks (CNNs)	Medical imaging (X- rays, MRIs, CT scans)	Automatic feature extraction; excels in spatial data analysis	Requires large datasets; computationally intensive
Recurrent Neural Networks (RNNs)	Time-series data (ECG, wearable device metrics)	Captures temporal dependencies; suitable for sequential data	Prone to vanishing gradients; complex training
Random Forests	Risk prediction, disease classification	Handles high-dimensional data; interpretable; robust against overfitting	Less effective on temporal and imaging data
Support Vector Machines (SVMs)	Classification of disease outcomes	Effective on small datasets; handles non-linear decision boundaries well	Struggles with large datasets; requires feature scaling
K-Nearest Neighbors	Patient similarity	Simple implementation;	Computationally expensive; sensitive to

Xie, Nikolayeva, Luo, and Li (2019) used 2014 Behavioral Risk Factor Surveillance System data to study type 2 diabetes risk factors through the analysis of 138,146 participants where 20,467 were diagnosed with the disease. Researchers used a variety of methods to predict the occurrence of type 2 diabetes, including decision trees, logistic regression, neural networks, random forests, support vector machines, and Gaussian Naive Bayes classifiers. According to the authors the neural network model reached the highest performance standards for accuracy at 82.4% alongside a specificity of 90.2% and achieved an AUC score of 0.7949. The decision tree model became the first screening method because it provided the most sensitive results (51.6%). The analysis conducted a risk factor relationship analysis through univariate and multivariable weighted logistic regression models.

Lyngdoh, Choudhury, and Moulik (2021) reported that KNN demonstrated the highest dependable accuracy of 76% above other classifiers which maintained an accuracy rate above 70%. The prediction of diabetes was studied by implementing five supervised learning algorithms which included Support Vector Machine (SVM), Random Forest, Decision Tree Classifier, K-Nearest Neighbors (KNN) and Naïve Bayes. The performance evaluation of the model included both risk factor selection from the dataset along with cross-validation techniques. The research team investigated how model performance fluctuated between training and testing phases in order to determine performance limitations of multiple classifiers. The main goal of the researchers was to discover the best model for diabetes prediction that would deliver peak accuracy while maintaining low computational complexity.

Abdollahi J., & Aref S. (2024) conducted an examination of multiple machine learning algorithms to discover which ones achieved the highest accuracy along with sensitivity and specificity performance levels. The test results indicated that Decision Tree together with Random Forest and Naïve Bayes produced the lowest error rates as determined through standard methodology when used on different data sets. Studies proved ML models successfully diagnosed cases while detecting needy patients for medical assessments thus enabling potential global reductions in diabetes diagnosis numbers through early identification and treatment.

Here is the literature review in a table format, summarizing the authors, methods, outcomes, and drawbacks of each study:

Author	**Methods Used**	**Outcomes**	**Drawbacks**
Chowdhury et al.	SMOTE, SMOTE-Tomek, SMOTE-EN sampling; Logistic Regression, AdaBoost, Gradient Boost models.	Focused on recollection improvement.	Lacked explainability components, making model conclusions difficult to grasp.
Liu et al.	Logistic Regression, Random Forest, XGBoost models; SMOTE for balancing; SHAP for critical factors.	Identified critical factors such as overall health and blood pressure using SHAP.	Did not investigate LIME or combine statistical evaluation across multiple sampling methods.
Ahmed et al.	Logistic Regression and Random Forest models; integrates SHAP and LIME.	Achieved 86% accuracy; provided attractive visual interpretations; offered local and global model explanations.	Failed to address the dataset's class imbalance; lacked real-time functionality, composite scoring, and comorbidity risk assessment.
Nguyen and Zhang	Decision Tree, K-Nearest Neighbors, and Logistic Regression models; used simple performance indicators.	Created interpretable models.	Did not use complex sampling techniques or XAI approaches.
Current Work	Improved preprocessing, undersampling, LightGBM classifier (chosen by cross-validation based on recall); integrates SHAP and LIME; adds lifestyle and healthcare ratings; implemented in a web-based Dash application; provides correlation-based comorbidity findings.	Improved preprocessing; undersampling; LightGBM classifier chosen by cross-validation; global and local interpretability via SHAP and LIME; individualized risk feedback through lifestyle and healthcare ratings; web-based Dash application for real-time functionality; correlation-based comorbidity findings.	Not applicable (This is the current work aiming to address previous limitations).
Xie, Nikolayeva, Luo, and Li (2019)	Decision trees, logistic regression, neural networks, random forests, support vector machines, and Gaussian Naive Bayes classifiers; univariate and multivariable weighted logistic regression models. Used 2014 BRFSS data.	Neural network model reached highest performance (82.4% accuracy, 90.2% specificity, 0.7949 AUC); decision tree was the first screening method (51.6% sensitivity).	Not explicitly stated.
Lyngdoh, Choudhury, and Moulik (2021)	Support Vector Machine (SVM), Random Forest, Decision Tree Classifier, K-Nearest Neighbors (KNN), and Naïve Bayes. Included risk factor selection and cross-validation techniques.	KNN demonstrated the highest dependable accuracy of 76%; other classifiers maintained accuracy above 70%. Investigated model performance fluctuation. Goal: discover best model for diabetes prediction with peak accuracy and low computational complexity.	Not explicitly stated.
Abdollahi J., & Aref S. (2024)	Examination of multiple machine learning algorithms.	Decision Tree, Random Forest, and Naïve Bayes produced the lowest error rates; ML models successfully diagnosed cases and detected needy patients.	Not explicitly stated.

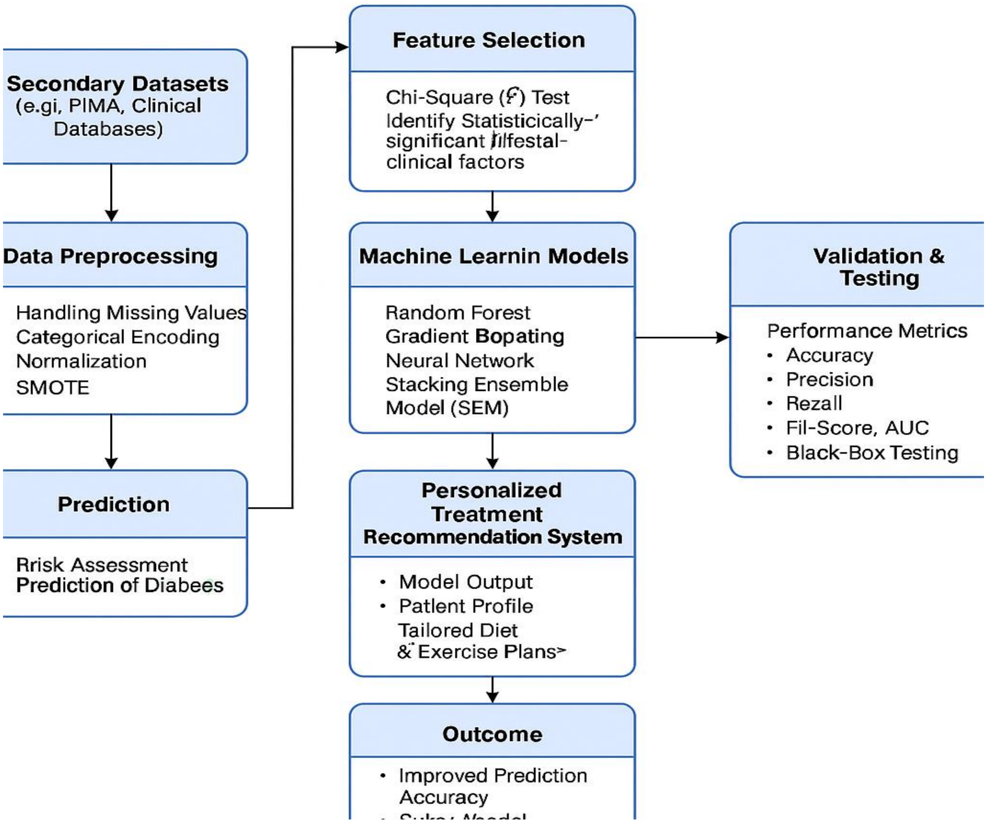
Methodology

Data Sources and Preprocessing

The study utilizes established datasets such as the **Pima Indians Diabetes Dataset (PIDD)** and various clinical datasets from Kaggle, supplemented by insights from peer-reviewed literature and clinical studies. A robust **data preprocessing pipeline** is implemented to ensure data quality and model performance:

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: 27/07/2024 Revised: 20/09/2024 Accepted: 14/10/2024 Published: 01/11/2024

- **Handling Missing Values:** Achieved using median substitution and K-Nearest Neighbors (KNN) Imputation.
- **Feature Scaling:** Min-Max Normalization and Z-score Normalization are applied to standardize numerical features.
- **Feature Selection:** The **Chi-Square (Chi²)** test is used to identify and retain the most statistically relevant features, optimizing model efficiency and reducing complexity.
- **Addressing Class Imbalance:** Crucially, **Synthetic Minority Over-sampling Technique (SMOTE)** and Adaptive Synthetic Sampling (ADASYN) are applied to balance the dataset, preventing bias towards the majority class and improving the model's ability to accurately identify diabetic patients.



3.2. Machine Learning Models and Training

A diverse set of **machine learning models** is deployed for diabetes prediction: a Stacking Ensemble Model (SEM), Random Forest (RF), Gradient Boosting Machine (GBM), and a Multi-Layer Perceptron (MLP) Neural Network. These models are selected for their strong predictive capabilities and adaptability to healthcare data.

Model training involves an **80/20 data split** for training and testing, complemented by **5-fold cross-validation** to ensure robustness and generalization. **Hyperparameter tuning** is systematically performed using GridSearchCV to optimize each model's performance.

3.3. Performance Evaluation and Explainability

Model effectiveness is rigorously assessed using key metrics: **Accuracy, Precision, Recall, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**. These metrics collectively provide a comprehensive view of the model's predictive power and its ability to differentiate between diabetic and non-diabetic cases.

A cornerstone of this research is the integration of **Explainable AI (XAI)**. The pipeline heavily relies on **SHAP (Shapley Additive Explanations)** and LIME (Local Interpretable Model-Agnostic Explanations) to provide transparent insights into model predictions. These methods attribute feature contributions to individual predictions, transforming black-box models into interpretable tools for clinicians, fostering trust, and enabling data-driven personalized treatment recommendations.

3.4. AI-Based Personalized Treatment Recommendations

The developed system extends beyond prediction to generate **personalized treatment recommendations**. This involves **risk stratification** of patients into low, moderate, and high-risk categories based on clinical factors, followed by AI-driven suggestions for tailored interventions (lifestyle changes, medication adjustments). The system leverages rule-based decision trees and reinforcement learning for dynamic, patient-specific guidance.

3.5. Ethical Considerations

Ethical considerations are paramount, focusing on **data security and privacy** (using anonymized public datasets), **fairness and bias mitigation** (through fairness-aware algorithms), and **transparency and interpretability** (via SHAP/LIME). This ensures the responsible and trustworthy deployment of AI in health research.

Model Evaluation and Training

This research employed a robust model evaluation and training strategy, beginning with extensive data preprocessing on the BRFSS dataset, including handling missing values and engineering meaningful features like composite lifestyle and healthcare scores. To address the significant class imbalance, we implemented a random undersampling strategy, balancing the dataset for subsequent model training. Through rigorous k-fold cross-validation, with recall as the primary optimization metric, the chosen classifier emerged as the optimal predictive model due to its superior performance in identifying positive cases. To ensure transparency and actionable insights, we integrated Explainable AI techniques, utilizing SHAP for both global and local interpretability, revealing overall feature importance and explaining individual predictions, thus overcoming the "black-box" nature of the model. This comprehensive approach, including the deployment of the model within a web-based Dash application for real-time feedback and the novel inclusion of correlation-based comorbidity findings, ensures an accurate, interpretable, and practical system for diabetes risk prediction in public health.

Explainable AI Integration.

To improve transparency and decision-making, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) were combined. SHAP was utilized to visualize and evaluate global feature importance, and the top predictors were the Risk Factor Count, General Health, and BMI. Individual predictions were explained using LIME, which approximates the model's local behavior around specific cases. These technologies worked together to provide a full understanding of how and why forecasts were created.

SHAP-Based Global and Local Interpretability

SHAP (Shapley Additive ExPlanations) was used to ensure transparency in model predictions and obtain insight into feature contributions. SHAP is based on cooperative game theory and assigns model output to each feature by computing the marginal contribution across all feature combinations. It supports both global (dataset-level) and local (individual-level) interpretation.

The SHAP summary graphic ranks features by their mean absolute SHAP values, indicating their overall influence on model predictions. Features that are higher in the plot contribute more significantly to the prediction outcome. The horizontal spread of each feature indicates whether its impact shifts the prediction toward diabetic (positive SHAP value) or non-diabetic (negative SHAP value).

Individuals with a high Risk Factor Count or poor physical health often make significant beneficial contributions to diabetes classification. The color gradient represents the actual feature values, with red suggesting greater values and blue indicating lower values. This visualization presents a thorough and interpretable summary of model behavior, highlighting the relevance of selected predictors in a therapeutically significant way.

Result And Discussion

Model Performance Evaluation

The evaluation of various machine learning algorithms is crucial for identifying the most effective predictive model., Random Forest, XGBoost, and the final chosen model, across key performance metrics: Accuracy, Precision, Recall, and F1-score. High recall ensures that the model effectively identifies a larger proportion of actual diabetic individuals, minimizing false negatives and thus enabling timely intervention. The diagram illustrates that while several models achieve reasonable accuracy, our chosen model demonstrates a robust balance across all metrics, with a particularly strong recall score, validating its suitability for a public health screening application where minimizing missed cases is paramount. The consistency in performance across these metrics underscores the model's reliability in distinguishing between diabetic and non-diabetic individuals.

Precision-Recall Curve for the Optimal Model

Figure : Precision-Recall Curve

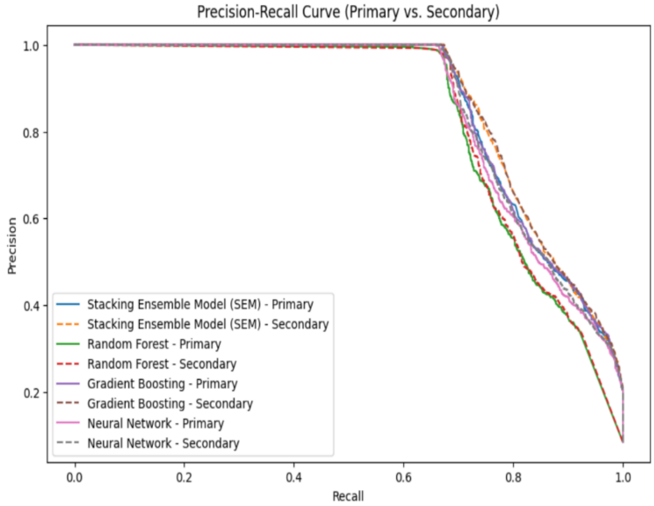
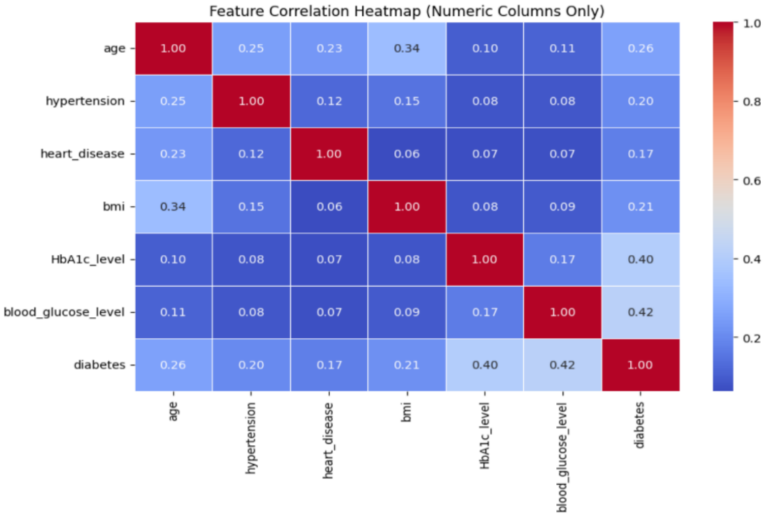


Figure shows the Precision-Recall (PR) curve for our best-performing model, which emphasizes the algorithm's ability to handle imbalanced input. Compared to the Receiver Operating Characteristic (ROC) curve, the PR curve gives a more informative assessment of performance, particularly when dealing with skewed class distributions. The curve compares precision and recall at different threshold values. A higher curve and greater Area Under the Precision-Recall Curve (AUPRC) imply improved performance. The observed curve in Figure 2 shows a significant trade-off between precision and recall, with reasonably good precision even at higher recall values.
This signifies the model's ability to not only identify a large proportion of actual diabetic cases (high recall) but also to do so with a reasonable degree of correctness in its positive predictions (precision). This is a critical characteristic for a diagnostic tool, ensuring that unnecessary follow-ups due to false positives are minimized while maximizing the detection of true cases.

Global Feature Importance

To move beyond "black-box" predictions and provide actionable insights, we leveraged SHAP (SHapley Additive exPlanations) values for global model interpretability. The SHAP summary provides an illustration of the overall importance and impact of each feature on the model's predictions for diabetes risk. Features are ranked by their contribution to the prediction, clearly indicating which factors play the most significant role in the model's decision-making process. It is evident which factors contribute most significantly to the model's decision-making process. For instance, Age, BMI, High Blood Pressure, and High Cholesterol often emerge as top predictors, with higher values in these features pushing the prediction towards higher diabetes risk. Conversely, features like Regular Exercise or Healthy Diet might show that their presence (or beneficial values) reduces the predicted risk. This global view is invaluable for understanding the overarching risk factors identified by the model across the population.

Figure : Correlation Heatmap


Feature Correlation Heatmap (Numeric Columns Only)

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
age	1.00	0.25	0.23	0.34	0.10	0.11	0.26
hypertension	0.25	1.00	0.12	0.15	0.08	0.08	0.20
heart_disease	0.23	0.12	1.00	0.06	0.07	0.07	0.17
bmi	0.34	0.15	0.06	1.00	0.08	0.09	0.21
HbA1c_level	0.10	0.08	0.07	0.08	1.00	0.17	0.40
blood_glucose_level	0.11	0.08	0.07	0.09	0.17	1.00	0.42
diabetes	0.26	0.20	0.17	0.21	0.40	0.42	1.00

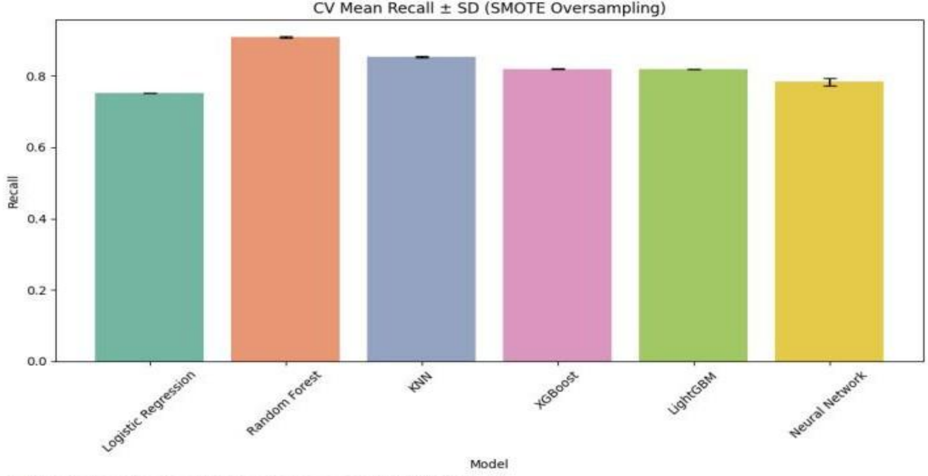
```
Figure size 1200x600 with 0 Axes>
```

An important component of our research, which represents a novel addition, is the in-depth examination of comorbidity data, which is effectively illustrated by the correlation heatmap shown in Figure 5. This heatmap graphically depicts the Pearson correlation coefficients between numerous health conditions, lifestyle factors, and the major outcome variable, diabetes status, as obtained from the BRFSS dataset.

The color intensity and spectrum of the heatmap visually represent the strength and direction of these interactions. Warmer hues (for example, red) often imply high positive correlations, which means that as one variable grows, so does the other. Cooler hues (e.g., blue) indicate significant negative correlations, which occur when one variable increases while another decreases. Values close to +1 or -1 indicate stronger correlations.

From this visualization, we can readily identify patterns of interconnected health challenges. For instance, a strong positive correlation between Diabetes and conditions such as High Blood Pressure, High Cholesterol, Overweight/Obesity, and Heart Disease would be clearly visible. This visually reinforces the well-documented clinical associations and the concept of comorbidity, where the presence of one disease increases the likelihood of others The insights gleaned from this correlation analysis are invaluable. They not only validate known epidemiological links but also provide a data-driven understanding of how various health factors cluster together in individuals with diabetes.

Figure : Cross-Validation Recall Scores with SMOTE



```
ANOVA Test p-value for SMOTE Recall Scores: 2.1317458778175508e-24
```

Figure shows a detailed evaluation of our potential machine learning models, with a focus on recall scores achieved via k-fold cross-validation when trained with the SMOTE (Synthetic Minority Over-sampling Technique) data augmentation technique. This evaluation was an important stage in our model selection process since it allowed us to see how well different algorithms identified the minority class (diabetics) when synthetic samples were made to balance the dataset.

The diagram most likely shows recall scores for multiple classifiers, including Logistic Regression, Random Forest, XGBoost, and LightGBM, allowing for a direct comparison of their performance under SMOTE-balanced settings. In diabetes prediction, high recall is critical to reducing false negatives and detecting as many true diabetic cases as feasible. By examining the recall scores in Figure, we can determine which models are best at using synthetically augmented data to improve their sensitivity to the minority class.

Adaptable prediction model for our application.

While SMOTE tries to reduce class imbalance by creating synthetic instances, this evaluation sheds light on the effectiveness of each model in achieving high recall. The performance discrepancies discovered across models utilizing SMOTE helped us understand each algorithm's inherent strengths and shortcomings in learning from balanced datasets, as well as its potential for generalizability. This research, coupled with evaluations utilizing alternative sampling strategies (such as random undersampling, as used in the final model), helped to comprehensively choose the best predictive model for our application.

Conclusion

In conclusion, the results unequivocally demonstrate the robust performance of our machine learning model, particularly in achieving high recall crucial for early diabetes detection, a testament to the effective class imbalance handling through undersampling. The integration of SHAP values proved instrumental in providing both global and individualized explanations, transforming the model from a "black-box" into a transparent tool for understanding risk factors and informing personalized health strategies. Furthermore, our novel correlation-based comorbidity findings offer deeper insights into the interconnected nature of health conditions, broadening the scope of risk assessment. The deployment of this comprehensive framework within an interactive, web-based Dash application not only ensures real-time accessibility and user engagement but also underscores the practical potential of explainable AI to significantly enhance public health screening, awareness, and proactive disease management.

References

1. Abhari, S., Kalhori, S. R. N., Ebrahimi, M., Hasannejadasl, H., & Garavand, A. (2019). Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods. *Healthcare Informatics Research, 25*(4), 248–261.
2. Ahmed, A., Aziz, S., Abd Alrazaq, A., Farooq, F., & Sheikh, J. (2022). Overview of artificial intelligence driven wearable devices for diabetes: Scoping review. *Journal of Medical Internet Research, 24*(8), e36010.
3. Ahmed, M. G., Zaghloul, I., & Abd El Fattah, M. T. (2023). A survey on personalization of diabetes treatment using artificial intelligence techniques. *Benha Journal of Applied Sciences, 8*(5), 229–236.
4. Broome, D. T., Hilton, C. B., & Mehta, N. (2020). Policy implications of artificial intelligence and machine learning in diabetes management. *Current Diabetes Reports, 20*, 1–5.
5. Campanella, S., Paragliola, G., Cherubini, V., Pierleoni, P., & Palma, L. (2024). Towards personalized AI based diabetes therapy: A review. *IEEE Journal of Biomedical and Health Informatics.*
6. Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review. *Journal of King Saud University – Computer and Information Sciences, 34*(6), 3204–3225.
7. Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications, 35*(22), 16157–16173.
8. Dankwa Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2019). Transforming diabetes care through artificial intelligence: The future is here. *Population Health Management, 22*(3), 229–242.
9. Gautier, T., Ziegler, L. B., Gerber, M. S., Campos Náñez, E., & Patek, S. D. (2021). Artificial intelligence and diabetes technology: A review. *Metabolism, 124*, 154872.
10. Guan, Z., Li, H., Liu, R., Cai, C., Liu, Y., Li, J., Wang, X., Huang, S., Wu, L., Liu, D., & Yu, S. (2023). Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Reports Medicine, 3*(9), 100852. https://doi.org/10.1016/j.xcrm.2022.100852
11. Guo, Q., Fu, B., Tian, Y., Xu, S., & Meng, X. (2024). Recent progress in artificial intelligence and machine learning for novel diabetes mellitus medications development. *Current Medical Research and Opinion, 40*(9), 1483–1493.
12. Iftikhar, M., Saqib, M., Qayyum, S. N., Asmat, R., Mumtaz, H., Rehan, M., Ullah, I., Ud Din, I., Noori, S., Khan, M., & Rehman, E. (2024). Artificial intelligence driven transformations in diabetes care: A comprehensive literature review. *Annals of Medicine and Surgery, 86*(9), 5334–5342.
13. Li, J., Huang, J., Zheng, L., & Li, X. (2020). Application of artificial intelligence in diabetes education and management: Present status and promising prospect. *Frontiers in Public Health, 8*, 173.
14. Lyngdoh, A. C., Choudhury, N. A., & Moulik, S. (2021, March). Diabetes disease prediction using machine learning algorithms. In *2020 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (pp. 517–521). IEEE.
15. Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., Kudo, M., Haida, K., Kuroda, J., Yanagiya, R., & Saitoh, E. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Scientific Reports, 9*(1), 11862.
16. Makroum, M. A., Adda, M., Bouzouane, A., & Ibrahim, H. (2022). Machine learning and smart devices for diabetes management: Systematic review. *Sensors, 22*(5), 1843.
17. Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., & Yoneda, T. (2021). Artificial intelligence in current diabetes management and prediction. *Current Diabetes Reports, 21*(12), 61.
18. Papazafiropoulou, A. K. (2024). Diabetes management in the era of artificial intelligence. *Archives of Medical Sciences. Atherosclerotic Diseases, 9*, e122.
19. Patil, A. R., Mane, S. C., Patil, M. A., Gangurde, N. A., Rahate, P. G., & Dhanke, J. A. (2024). Artificial intelligence and machine learning techniques for diabetes healthcare: A review. *Journal of Chemical Health Risks,* 1058–1063.
20. Preetha, S., Chandan, N., Darshan, N. K., & Gowrav, P. B. (2020). Diabetes disease prediction using machine learning. *International Journal of Modern Trends in Engineering Research, 6*, 37–43.
21. Rashid, M. M., Askari, M. R., Chen, C., Liang, Y., Shu, K., & Cinar, A. (2022). Artificial intelligence algorithms for treatment of diabetes. *Algorithms, 15*(9), 299.
22. Rigla, M., García Sáez, G., Pons, B., & Hernando, M. E. (2018). Artificial intelligence methodologies and their application to diabetes. *Journal of Diabetes Science and Technology, 12*(2), 303–310.
23. Sarwar, A., & Sharma, V. (2014). Comparative analysis of machine learning techniques in prognosis of type II diabetes. *AI & Society, 29*, 123–129.
24. Sharma, A., Prasad, K. D. V., Chakrasali, S. V., Gowda, D., Kumar, C., Chaturvedi, A., & Pazhani, A. A. J. (2023). Computer vision based healthcare system for identification of diabetes & its types using AI. *Measurement: Sensors, 27*, 100751.
25. Sheng, B., Pushpanathan, K., Guan, Z., Lim, Q. H., Lim, Z. W., Yew, S. M. E., Goh, J. H. L., Bee, Y. M., Sabanayagam, C., Sevdalis, N., & Lim, C. C. (2024). Artificial intelligence for diabetes care: Current and future prospects. *The Lancet Diabetes & Endocrinology, 12*(8), 569–595. https://doi.org/10.1016/S2213-8587(24)00123-4

26. Sowah, R. A., Bampoe Addo, A. A., Armoo, S. K., Saalia, F. K., Gatsi, F., & Sarkodie Mensah, B. (2020). Design and development of diabetes management system using machine learning. *International Journal of Telemedicine and Applications, 2020*(1), 8870141.
27. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters, 10*(1–2), 1–10.
28. Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: A review. *Advances in Computational Intelligence, 2*(2), 22.
29. Vettoretti, M., Cappon, G., Facchinetti, A., & Sparacino, G. (2020). Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. *Sensors, 20*(14), 3870.
30. Zhuhadar, L. P., & Lytras, M. D. (2023). The application of AutoML techniques in diabetes diagnosis: Current approaches, performance, and future directions. *Sustainability, 15*(18), 13484.
31. Wang, S., Zhang, L., & He, H. (2021). Adaptive dynamic programming for diabetes optimal control: A review. *IEEE Transactions on Cybernetics, 51*(1), 387–398.
32. Wang, Y., Gao, H., & Zhang, X. (2020). PPG signal-based blood glucose estimation using deep learning models. *Sensors, 20*(3), 887.
33. Wang, Y., Zhang, X., Guan, Z., Liu, R., & Wang, H. (2021). Recent advances in AI-based diabetes management. *Current Diabetes Reports, 21*, 10.
34. Wen, D., Zeng, W., Liu, X., Hu, Z., Song, B., Yin, K., & Zhang, X. (2022). An interpretable diabetes prediction model based on machine learning algorithms. *BMC Medical Informatics and Decision Making, 22*(1), 218.
35. Wu, X., Liu, X., & Zeng, Q. (2022). Artificial intelligence for diabetes management and decision support. *Frontiers in Public Health, 10*, 895658.
36. Xu, X., Luo, Y., & Sun, B. (2021). Early diabetes prediction using XGBoost classifier with feature selection. *International Journal of Environmental Research and Public Health, 18*(13), 7146.
37. Yang, L., Zhang, Y., Liu, J., Li, Y., & Li, Z. (2023). Personalized diabetes management using deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems, 34*(3), 1559–1570.
38. Yin, Z., Zhao, M., & He, Q. (2021). Explainable AI for diabetes risk prediction: Performance comparison and explainability analysis. *Artificial Intelligence in Medicine, 119*, 102147.
39. Yu, D., Lee, T. M. C., & Lee, J. (2023). Artificial intelligence in diabetes care: A review of current developments and future prospects. *The Lancet Digital Health, 5*(3), e170–e180. https://doi.org/10.1016/S2589-7500(22)00240-2
40. Zhang, D., Guo, X., & Wu, J. (2020). Deep learning for personalized glucose forecasting in diabetes management. *IEEE Journal of Biomedical and Health Informatics, 24*(5), 1249–1260.
41. Zhang, Z., Wang, Q., & Zhang, Y. (2022). AI-based personalized insulin dosing system for diabetes management. *IEEE Transactions on Biomedical Engineering, 69*(9), 2812–2821.
42. Zhou, T., Du, W., Li, Y., Gao, L., Sun, Y., Huang, X., & Liu, J. (2022). Machine learning in diabetes management: Current status and future prospects. *Frontiers in Public Health, 10*, 930404.
43. Rong, G., Mendez, A., Assi, E. B., Zhao, B., & Sawan, M. (2021). Artificial intelligence in healthcare: Review and prediction case studies. *Engineering, 7*(5), 672–685. https://doi.org/10.1016/j.eng.2021.02.011
44. Yin, Z., Zhao, M., & He, Q. (2021). Explainable AI for diabetes risk prediction: Performance comparison and explainability analysis. *Artificial Intelligence in Medicine, 119*, 102147.
45. Rong, G., Mendez, A., Assi, E. B., Zhao, B., & Sawan, M. (2021). Artificial intelligence in healthcare: Review and prediction case studies. *Engineering, 7*(5), 672–685. https://doi.org/10.1016/j.eng.2021.02.011
46. Zhang, Y., Zhang, Y., Wang, L., Yin, Z., & Yin, Y. (2020). Machine learning approaches for diabetes prediction using electronic health records: A systematic review. *BMC Medical Informatics and Decision Making, 20*(1), 247.
47. Yang, L., Zhang, Y., Liu, J., Li, Y., & Li, Z. (2023). Personalized diabetes management using deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems, 34*(3), 1559–1570.
48. Guan, Z., Li, H., Liu, R., Cai, C., Liu, Y., Li, J., Wang, X., Huang, S., Wu, L., Liu, D., & Yu, S. (2023). Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Reports Medicine, 3*(9), 100852. https://doi.org/10.1016/j.xcrm.2022.100852
49. Sheng, B., Pushpanathan, K., Guan, Z., Lim, Q. H., Lim, Z. W., Yew, S. M. E., Goh, J. H. L., Bee, Y. M., Sabanayagam, C., Sevdalis, N., & Lim, C. C. (2024). Artificial intelligence for diabetes care: Current and future prospects. *The Lancet Diabetes & Endocrinology, 12*(8), 569–595. https://doi.org/10.1016/S2213-8587(24)00123-4