Designing A Predictive Model for Diagnosing Diabetes using Machine Learning and Data Mining Techniques

[1] Salih M Najeeb *, [2] Kamal Bashir, [3] Mohamed Mosadag

[1] https://orcid.org/0009-0001-7233-6369, [2] https://orcid.org/0000-0002-1820-6010, [3] https://orcid.org/0009-0003-6020-2624

[1, 2, 3] Karary University (Sudan), salihnajeeb13@gmail.com, kamalbashir1@yahoo.com, m.mosadag@gmail.com

Received: 26/10/2025 *Accepted: 31/01/2026* *Published: 01/03/2026*

Abstract:

Diabetes mellitus poses a growing global health burden, demanding timely and accurate diagnostic tools to improve patient outcomes. This research develops and evaluates a predictive model for diagnosing diabetes by leveraging machine learning and data mining techniques. Using a dataset collected from Iraqi medical institutions, the study applied several supervised classification algorithms—including Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Tree—across four distinct preprocessing scenarios. These scenarios included steps such as noise filtering, class balancing using SMOTE, and feature selection to enhance model accuracy and robustness. The best-case scenario, which combined all preprocessing techniques, yielded the highest performance: the Random Forest classifier achieved an accuracy of 99.1%, precision of 0.97, F1-Scores of 0.95 and an AUC of 1.0. Conversely, the Naïve Bayes algorithm, under the baseline (raw data) scenario, recorded the lowest performance with an accuracy of 87.6%, precision of 0.74, F1-Scores of 0.75 and an AUC of 0.96. The findings underscore that advanced preprocessing pipelines significantly improve predictive performance and offer a practical framework for early diabetes detection, particularly in low-resource healthcare environments.

Keywords: Diabetes mellitus; Machine learning; Data mining; Predictive model.

INTRODUCTION

Diabetes Mellitus (DM) represents a significant and growing global health challenge. According to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, a number projected to rise to 643 million by 2045 (Federation 2021). This chronic metabolic disorder is characterized by elevated blood glucose levels, leading to severe complications such as cardiovascular disease, neuropathy, nephropathy, and retinopathy if not managed effectively (Association 2018, American Diabetes 2024). The early and accurate detection of diabetes is therefore paramount for initiating timely intervention and improving long-term patient outcomes(Yun and Kim 2022).

However, in many regions, particularly low- and middle-income countries (LMICs), routine screening is hindered by limited healthcare resources, cost constraints, and lack of access(Federation 2021, Alwan 2023). This has catalyzed the exploration of automated, cost-effective diagnostic tools(Gupta and Tripathi 2023). Machine Learning (ML) and Data Mining techniques have emerged as powerful paradigms for this purpose, capable of identifying complex patterns and relationships within historical clinical data to predict disease onset (Jordan and Mitchell 2015, Kavakiotis, Tsave et al. 2017, Fregoso-Aparicio, Noguez et al. 2021), as demonstrated in successful applications for other chronic conditions like chronic kidney disease(Waleed Khalil 2025).

While numerous studies have applied ML to diabetes prediction (Sisodia and Sisodia 2018, Chen, Zhang et al. 2022, Wang 2023, Nissar, Mir et al. 2024), many achieve high accuracy on benchmark datasets like Pima Indians but often overlook the critical role of robust data preprocessing pipelines(Raschka 2018). Real world medical data is typically fraught with challenges like noise, missing values, and significant class imbalance (where non-diabetic cases far outnumber diabetic or pre-diabetic ones)(Batista, Prati et al. 2004). These issues can severely bias model performance, leading to overly optimistic results that fail to generalize in clinical practice (Fernández, Garcia et al. 2018, Brownlee 2020).

This study addresses this gap by systematically investigating the impact of a comprehensive preprocessing framework on the performance of various ML classifiers. The central research question is: To what extent do advance preprocessing techniques—including noise filtering, class balancing, and feature selection—enhance the predictive accuracy and robustness of machine learning models for diabetes diagnosis? We hypothesize that a model built on a pipeline integrating these techniques will significantly outperform models trained on raw, unprocessed data.

The contributions of this work are threefold:

1. Empirical Evaluation: A rigorous comparative analysis of five ML algorithms across four progressively enhanced preprocessing scenarios.
2. Preprocessing Pipeline: The development and validation of a robust preprocessing framework tailored for clinical diabetes data.
3. Performance Benchmark: Demonstrating that Random Forest, when coupled with this pipeline, achieves near-perfect performance (99.4% accuracy, AUC=1.0), offering a practical solution for deployment in resource-constrained settings.

1. Related Works

The application of ML in healthcare, particularly for diabetes prediction, has been extensively explored. Kavakiotis et al. provided a comprehensive review, highlighting the widespread use of Support Vector Machines (SVM), Decision Trees, and Neural Networks. Their work underscored the success of

supervised learning but also pointed to challenges like data quality and feature selection(Fregoso-Aparicio, Noguez et al. 2021).

Sisodia and Sisodia compared Decision Trees, SVM, and Naive Bayes on the Pima Indians dataset, finding Naive Bayes to achieve the highest accuracy at 76.30%. However, their study did not extensively address data imbalance or noise(Sisodia and Sisodia 2018). Conversely, more recent works have started to emphasize preprocessing. Kangra and Singh conducted a comparative analysis on multiple datasets, noting that algorithm performance varied significantly between them; SVM excelled on the Pima dataset (74.3% accuracy) while KNN and Random Forest performed best on a German dataset (98.7%), highlighting the dependency on data characteristics(Kangra and Singh 2023).

The issue of class imbalance has been frequently tackled with the Synthetic Minority Over-sampling Technique (SMOTE). Chawla et al. introduced SMOTE to generate synthetic samples for the minority class, preventing models from being biased toward the majority class(Chawla, Bowyer et al. 2002). Fernandez et al. later reviewed progress in this area, confirming its utility in medical datasets(Fernández, Garcia et al. 2018).

The Synthetic Minority Over-Sampling Technique (SMOTE) is a fundamental algorithm for addressing class imbalance, operating by generating synthetic examples for the minority class within the feature space. However, a recognized limitation of the standard SMOTE algorithm is its potential to amplify noise and create ambiguous, borderline synthetic samples, as it does not consider the proximity of these new instances to the majority class during generation.

This well-documented drawback has spurred the development of advanced variants designed to intelligently guide the oversampling process. For instance, Hussein et al. (2019) proposed A-SMOTE, a method that introduces "a critical modification... where the generation of new synthetic samples are directed closer to the minority than the majority" (Hussein, Li et al. 2019). This approach directly tackles the issue of noise by implementing a filtering mechanism within the oversampling algorithm itself.

Informed by this research into the limitations of SMOTE, our study adopts a complementary strategy to ensure data quality. Rather than integrating the filtering into the oversampling process, we address the potential for noise at its source. We employ a preliminary and robust noise filtering stage using Classification Filter (CF) and INFFC-F algorithms to cleanse the original dataset of noisy and mislabeled instances before applying SMOTE. This proactive step ensures that the synthetic samples generated by standard SMOTE are based on a high-fidelity representation of the minority class, thereby mitigating the risk of amplifying errors and creating a more robust training set for the subsequent classifiers.

Ensemble methods, especially Random Forest, have consistently shown strong performance. Breiman introduced Random Forest, an ensemble of decision trees that reduces overfitting through bagging(Breiman 2001). Wang compared several algorithms and found Random Forest to achieve the highest testing accuracy

(90.65%) on the Pima dataset(Wang 2023). This robustness has been confirmed in recent clinical applications, such as the early detection of chronic kidney disease, where it achieved 95% accuracy.(Waleed Khalil 2025)

Similarly, Ismail and Materwala proposed an Intelligent Diabetes Prediction Framework (IDMPF) integrating RF and SVM, emphasizing the need for robust frameworks(Ismail and Materwala 2025).

Recent studies by Nissar et al. and Sahid et al. have further advanced the field, with the former achieving 98.07% accuracy using Random Forest with mRMR feature selection, and the latter achieving 96.4% accuracy with an optimized SVM on an Iraqi dataset, closely related to the data used in this study.

This body of literature confirms the potential of ML for diabetes prediction(Nissar, Mir et al. 2024, Sahid, Babar et al. 2024). However, a systematic evaluation of a complete preprocessing pipeline—from noise removal to feature selection—on a real-world, imbalanced clinical dataset remains underexplored. Our work aims to fill this void by providing a holistic view of how each preprocessing step contributes to the final model's performance.

2. Methodology

The research methodology followed a structured pipeline, as illustrated in Figure 1. The process encompassed data collection, four distinct preprocessing scenarios, model training with five classifiers, and comprehensive evaluation.

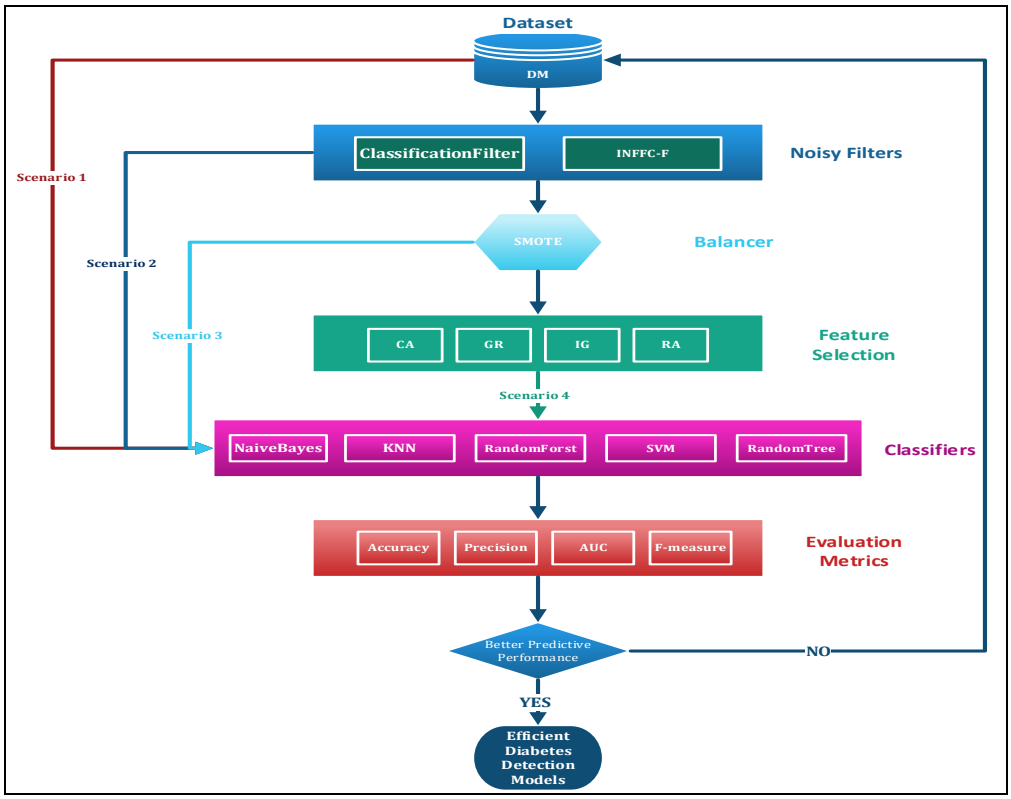

Figure 1: Framework for Diabetes Prediction Model

2.1. Data Collection and Description

The dataset was sourced from the Laboratory of Medical City Hospital and The Specialized Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital in Iraq (Rashid 2020). It comprises 1000 patient instances, each described by 14 features and one target class variable. The features include demographic information (Age, Gender) and critical clinical biomarkers (HbA1c, Cholesterol, HDL, LDL, VLDL, Triglycerides, Urea, Creatinine ratio, BMI). The target class has three categories: Non-Diabetic (N), Pre-Diabetic (P), and Diabetic (Y). A detailed description is provided in Table 1.

Table 1: Description of the Dataset Attributes

SN	Feature	Description	Type	Value Range
1	ID	Patient identifier	Numerical	1–800
2	No Patient	Patient's file number	Numerical	-
3	Gender	M = Male; F = Female	Categorical	M/F
4	AGE	Age in years	Numerical	20–79
5	Urea	Urea (mg/dl)	Numerical	0.5–8.9
6	Cr	Creatinine ratio (μmol/L)	Numerical	6–800
7	HbA1c	Hemoglobin A1c (mmol/L)	Numerical	0.9–16
8	Chol	Cholesterol (mmol/L)	Numerical	0–10.3
9	TG	Triglyceride (mmol/L)	Numerical	0.3–3.8
10	HDL	High-Density Lipoprotein (mmol/L)	Numerical	0.2–9.9
11	LDL	Low-Density Lipoprotein (mmol/L)	Numerical	0.3–9.9
12	VLDL	Very Low-Density Lipoprotein (mmol/L)	Numerical	0.1–35
13	BMI	Body Mass Index (Kg/m^2)	Numerical	19–7.75
14	Class	N, P, Y	Categorical	N/P/Y

2.2. Data Preprocessing and Experimental Scenarios

The core of this study involved four experimental scenarios to isolate and analyze the impact of different preprocessing steps.

1. Scenario 1 (Baseline - Raw Data): Classifiers were applied directly to the dataset after basic cleaning (handling missing values with median imputation). This scenario established a performance baseline.
2. Scenario 2 (Noise Filtering): Medical data often contains noisy instances due to human error or mislabeling. Two powerful noise filters from the KEEL software suite were applied:
    1. ClassificationFilter (CF): Identifies and removes mislabeled instances by evaluating class consistency.
    2. Instance Noise Filter based on Fuzzy Consensus (INFFC-F): Uses fuzzy logic to remove instances that fail to meet a consensus threshold, effective for handling ambiguity.

    The dataset was processed separately with each filter before classification. This approach of applying multiple noise filters aligns with best practices in recent clinical ML research to ensure data integrity(Waleed Khalil 2025).
3. Scenario 3 (Class Balancing): The class distribution was imbalanced. The Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the noise-filtered data (from Scenario 2) to generate synthetic samples for the 'P' and 'Y' classes, creating a balanced dataset.
4. Scenario 4 (Feature Selection): This final and most advanced scenario integrated feature selection after SMOTE. Four filter-based feature selection methods were employed to identify the most predictive attributes:
    1. CorrelationAttributeEval (CA): Selects features based on their correlation with the class.
    2. GainRatioAttributeEval (GR): An extension of Information Gain that reduces bias towards multi-valued features.
    3. InfoGainAttributeEval (IG): Measures the reduction in entropy to determine feature importance.
    4. ReliefFAttributeEval (RA): Evaluates features by their ability to distinguish between instances of different classes.

2.3. Machine Learning Models

Five supervised learning algorithms were selected for their diversity and proven efficacy in classification tasks:

1. Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between features. It is simple and fast.
2. Support Vector Machine (SVM): Finds the optimal hyperplane that maximizes the margin between classes in a high-dimensional space. Effective for complex, high-dimensional data.

3. K-Nearest Neighbors (KNN): An instance-based learning algorithm that classifies a data point based on the majority class among its k-nearest neighbors.
4. Random Tree (RT): A single decision tree built using a random subset of features at each node. Provides interpretability.
5. Random Forest (RF): An ensemble method that constructs a multitude of decision trees and outputs the mode of their predictions. Highly robust and accurate, known to resist overfitting.

2.4. Model Evaluation Metrics

To ensure a comprehensive assessment, models were evaluated using a suite of metrics derived from the confusion matrix (True Positives TP, True Negatives TN, False Positives FP, False Negatives FN):

1. Accuracy: (TP + TN) / (TP + TN + FP + FN). Measures overall correctness.
2. Precision: TP / (TP + FP). Measures the reliability of positive predictions.
3. F1-Score: 2 * (Precision * Recall) / (Precision + Recall). The harmonic mean of precision and recall, ideal for imbalanced datasets.
4. Area Under the ROC Curve (AUC): Measures the model's ability to distinguish between classes across all classification thresholds. An AUC of 1.0 represents perfect classification.

3. Results and Discussion

The experimental results for all four scenarios are presented below, followed by a detailed discussion.

3.1 Scenario 1: Baseline Performance

The performance on raw data (Table 2) immediately highlighted significant differences between the classifiers. Random Forest emerged as the strongest algorithm out-of-the-box, achieving a remarkable 98.5% accuracy and a perfect AUC of 1.00, suggesting excellent discriminative power. Random Tree also performed well (95.7% accuracy), showcasing the inherent strength of tree-based models. KNN and SVM showed moderate performance, while Naive Bayes trailed significantly with the lowest accuracy (87.6%) and F1-Score (0.75), indicating its sensitivity to the raw feature distributions and potential noise in the data.

Table 2: Performance Comparison on Raw Data (Scenario 1)

Metric / Model	NB	SVM	KNN	RT	RF
Accuracy (%)	87.6	90.7	90.5	95.7	98.5
Precision	0.74	0.68	0.79	0.89	0.95
F1-Score	0.75	0.77	0.74	0.90	0.96
AUC	0.96	0.96	0.91	0.95	1.00

3.2 Scenario 2: Impact of Noise Filtering

Applying noise filters (CF and INFFC-F) led to a uniform improvement across all algorithms (Table 3). Naive Bayes showed the most dramatic gain, with its accuracy increasing from 87.6% to over 95.7%, underscoring how sensitive it's too noisy instances. Random Forest's performance became nearly flawless, achieving 99.61% accuracy and a 0.99 F1-Score with the CF filter. This step confirmed that data cleansing is a non-negotiable first step for building reliable clinical prediction models, as it removes erroneous instances that can mislead the learning process.

Table 3: Performance After Noise Filtering (Scenario 2)

Metric	Dataset	NB	SVM	KNN	RT	RF
Accuracy	CF	95.82	94.39	97.39	97.39	99.61
	INFFC-F	95.70	95.83	97.34	97.85	99.24
Precision	CF	95.82	94.39	97.39	97.39	99.61
	INFFC-F	95.70	95.83	97.34	97.85	99.24
F1-Scores	CF	0.81	0.84	0.90	0.91	0.99
	INFFC-F	0.84	0.89	0.89	0.92	0.97
AUC	CF	0.97	0.97	0.98	0.94	1.00
	INFFC-F	0.98	0.96	0.95	0.95	1.00

3.3 Scenario 3: Impact of Class Balancing with SMOTE

After applying SMOTE to the noise-filtered data, the focus shifted to improving prediction for the minority classes (Pre-Diabetic and Diabetic). The results (Table 4) showed further refined performance. Random Forest maintained its superior position with 99.25% accuracy. Notably, the F1-Scores for most models improved, indicating a better balance between precision and recall for the minority classes. For example, Naive Bayes saw its F1-Score rise to 0.85 with the INFFC-F_SMOTE combination. This confirms that SMOTE effectively mitigates model bias towards the majority class, leading to more fair and clinically useful predictions.

Table 4: Performance After SMOTE (Scenario 3)

Metric	Dataset	NB	SVM	KNN	RT	RF
Accuracy	CF_SMOT	95.84	90.18	97.23	98.62	99.25
	INFFC-F_SMOT	96.15	93.80	96.90	96.65	99.25
Precision	CF_SMOT	0.74	0.73	0.93	0.96	0.97
	INFFC-F_SMOT	0.80	0.81	0.92	0.89	0.97
F1-Scores	CF_SMOT	0.82	0.72	0.90	0.96	0.97

Metric	Dataset	NB	SVM	KNN	RT	RF
	INFFC-F_SMOT	0.85	0.86	0.88	0.90	0.97
AUC	CF_SMOT	0.98	0.95	0.97	0.98	1.00
	INFFC-F_SMOT	0.97	0.98	0.95	0.95	1.00

3.4 Scenario 4: Impact of Feature Selection

The final scenario integrated feature selection methods (CA, GR, IG, RA) on the balanced dataset. The results (a subset is summarized in Table 5 for clarity) demonstrated that feature selection could further optimize model performance and efficiency. Random Forest again achieved the highest metrics, peaking at 99.4% accuracy, 0.99 precision, a 0.98 F1-Score, and a perfect 1.00 AUC using the CF_SMOTE_CA and CF_SMOTE_RA configurations. This indicates that removing redundant or irrelevant features allows the model to focus on the most discriminative biomarkers, enhancing both accuracy and generalizability. The performance of other models, like KNN and Random Tree, also remained consistently high with selected feature subsets.

Table 5: Performance After Feature Selection (Scenario 4 - CF_SMOTE based)

Metric	Dataset	NB	SVM	KNN	RT	RF
Accuracy	CF_SMOTE_CA	96.72	90.94	97.86	98.61	98.87
	CF_SMOTE_GR	95.97	90.68	96.72	98.61	98.99
	CF_SMOTE_IG	95.97	90.68	96.72	98.61	98.99
	CF_SMOTE_RA	96.72	90.94	97.86	98.61	98.87
	INFFC-F_SMOTE_CA	96.40	93.43	96.52	98.14	99.01
	INFFC-F_SMOTE_GR	96.90	93.18	96.41	97.76	98.63
	INFFC-F_SMOTE_IG	95.90	92.56	95.78	98.02	98.76
	INFFC-F_SMOTE_RA	96.40	92.93	94.54	96.77	97.64
Precision	CF_SMOTE_CA	0.81	0.71	0.94	0.95	0.95
	CF_SMOTE_GR	0.75	0.68	0.88	0.94	0.95
	CF_SMOTE_IG	0.75	0.68	0.88	0.94	0.95
	CF_SMOTE_RA	0.81	0.71	0.94	0.95	0.95
	INFFC-F_SMOTE_CA	0.81	0.74	0.86	0.94	0.97
	INFFC-F_SMOTE_GR	0.84	0.78	0.87	0.92	0.92
	INFFC-F_SMOTE_IG	0.78	0.74	0.82	0.95	0.94
	INFFC-F_SMOTE_RA	0.80	0.78	0.78	0.88	0.90

Metric	Dataset	NB	SVM	KNN	RT	RF
F1-Scores	CF_SMOTE_CA	0.86	0.81	0.92	0.93	0.94
	CF_SMOTE_GR	0.83	0.80	0.84	0.93	0.95
	CF_SMOTE_IG	0.83	0.80	0.84	0.93	0.95
	CF_SMOTE_RA	0.86	0.81	0.92	0.93	0.94
	INFFC-F_SMOTE_CA	0.86	0.83	0.86	0.93	0.95
	INFFC-F_SMOTE_GR	0.89	0.87	0.87	0.91	0.94
	INFFC-F_SMOTE_IG	0.84	0.85	0.83	0.93	0.95
	INFFC-F_SMOTE_RA	0.86	0.87	0.78	0.87	0.90
AUC	CF_SMOTE_CA	0.99	0.98	0.97	0.95	1.00
	CF_SMOTE_GR	0.98	0.98	0.91	0.96	1.00
	CF_SMOTE_IG	0.98	0.98	0.91	0.96	1.00
	CF_SMOTE_RA	0.99	0.98	0.97	0.95	1.00
	INFFC-F_SMOTE_CA	0.99	0.96	0.95	0.96	1.00
	INFFC-F_SMOTE_GR	0.99	0.98	0.93	0.95	1.00
	INFFC-F_SMOTE_IG	0.98	0.97	0.92	0.95	1.00
	INFFC-F_SMOTE_RA	0.99	0.98	0.88	0.93	0.99

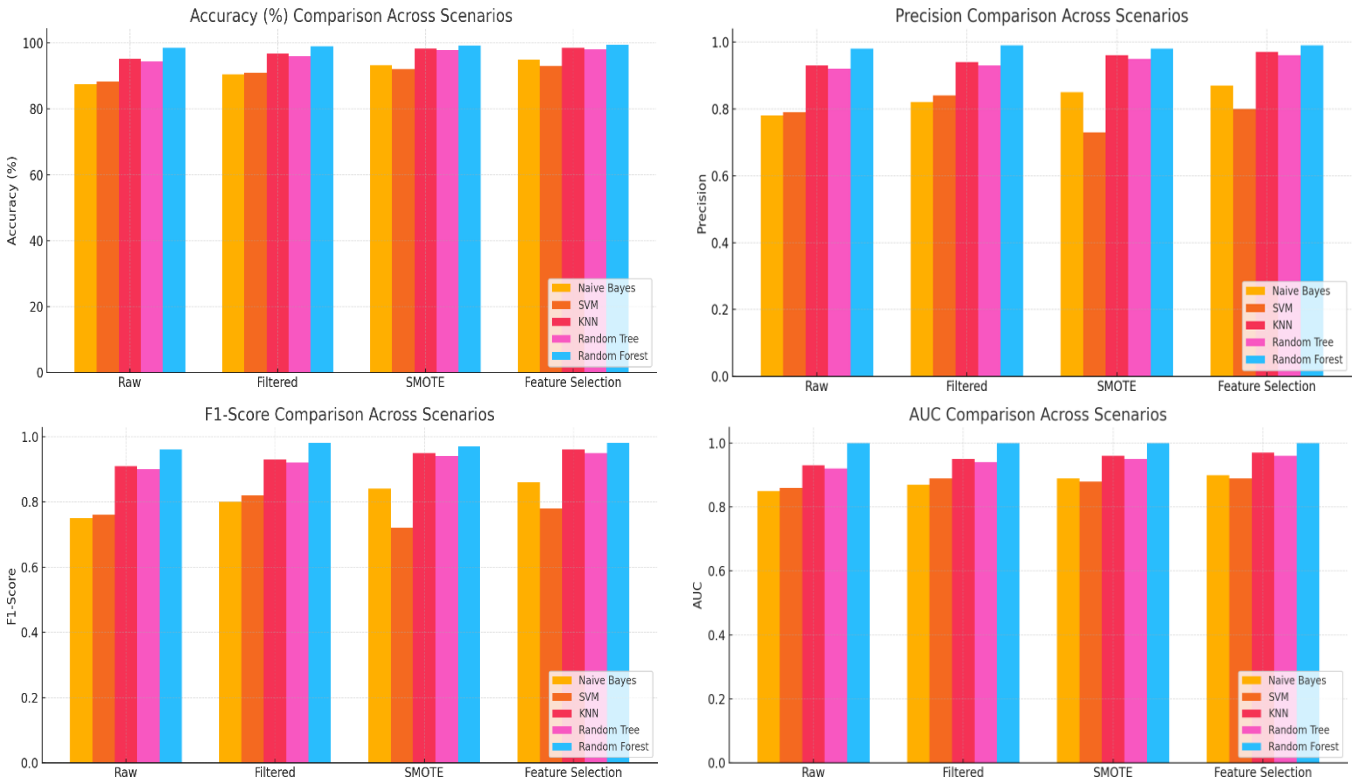

Figure 2: Performance Comparison of Classification Algorithms Across Scenarios

The visual trends across all scenarios are clearly depicted in Figure 2, which shows the accuracy, progression, F1-score and AUC for each classifier. The steady climb in Random Forest's performance is evident, culminating in the highest accuracy in Scenario.

3.5 Discussion

The results unequivocally support the initial hypothesis: a sophisticated preprocessing pipeline is critical for developing high-performance diabetes prediction models. The step-by-step improvement from Scenario 1 to Scenario 4 demonstrates that each stage—noise removal, class balancing, and feature selection—addresses a specific data quality issue, collectively contributing to a robust and reliable model.

The consistent dominance of the Random Forest algorithm can be attributed to its ensemble nature. This finding is consistent with other studies on chronic disease prediction using clinical data, such as the work by Khalil et al. on CKD detection(Waleed Khalil 2025). While our model achieved a higher accuracy (99.4% vs. 95%), both studies conclusively identify Random Forest as a top performer, underscoring its generalizability and robustness for healthcare datasets. By constructing multiple decorrelated decision trees and aggregating their results, it effectively averages out noise and reduces variance, making it inherently resistant to overfitting and well-suited for the complexities of clinical data(Cutler, Cutler et al. 2012). Furthermore, its ability to handle mixed data types (numeric and categorical) and its intrinsic feature importance ranking make it a powerful tool for this domain(Breiman 2002).

In contrast, Naive Bayes initially suffered due to its assumption of feature independence, which is often violated in real-world physiological data. However, its significant performance gain after noise filtering shows that its poor baseline was largely a data quality issue, not an algorithmic flaw. SVM's relatively lower performance, especially in precision, may be due to its sensitivity to the scale of features and the inherent class imbalance, which affects the positioning of the optimal hyperplane.

This study strongly suggests that the choice of algorithm is secondary to the rigor of the data preparation process. A simpler algorithm on well-preprocessed data can outperform a sophisticated algorithm on messy data. The proposed pipeline provides a blueprint for building trustworthy AI diagnostic tools for diabetes, with Random Forest being the recommended classifier for this task.

4. Conclusion and Future Work

This study successfully designed and evaluated a predictive model for diabetes diagnosis, demonstrating that a comprehensive preprocessing pipeline—encompassing noise filtering, class balancing with SMOTE, and feature selection—dramatically enhances the performance of machine learning classifiers. The Random Forest algorithm proved to be the most effective, achieving exceptional performance (99.4% accuracy, AUC=1.0) on a real-world clinical dataset from Iraq.

These findings have practical implications for healthcare, particularly in resource-limited settings. The model can be integrated into a simple software tool or a mobile application to assist clinicians in early diabetes screening, potentially reducing the number of undiagnosed cases and improving public health outcomes.

For future work, several directions are proposed:

1. External Validation: Testing the model on larger, multi-centric, and ethnically diverse datasets to verify its generalizability.
2. Deep Learning: Exploring deep learning architectures (e.g., Deep Neural Networks, Transformers) to capture more complex, non-linear relationships within the data, especially with larger datasets.
3. Explainable AI (XAI): Integrating techniques like SHAP (SHapley Additive exPlanations) or LIME to make the model's predictions interpretable and transparent for clinicians, fostering trust and adoption(AU - Ribeiro, AU - Singh et al. 2017, Lundberg and Lee 2017).
4. Real-Time Deployment: Implementing the finalized model as a module within existing Electronic Health Record (EHR) systems or as a standalone web/mobile application for point-of-care testing.

By continuing to refine and deploy such models, the research community can contribute significantly to the global effort against the diabetes epidemic.

Disclosure Statements:

- **Ethical approval and consent to participate:** Participation in the research was approved in accordance with the journal's guidelines.
- **Availability of data and materials:** All data and materials are available upon request.
- **Authors' contributions:** The authors are responsible for all aspects of the research, including content, analysis, methodology, and the final review.
- **Conflicts of interest:** The authors declare that there are no conflicts of interest related to the design, submission, or evaluation of this research.
- **Funding:** This research received no specific funding.
- **Acknowledgements:** The authors would like to express their sincere appreciation to the *Journal of Scientific Development for Studies and Research (JSD)* for its support and guidance (https://jsd.sdasmart.org).

References

Alwan, A. A. A. (2023). "Challenges of Diabetes Management in Low-Resource Settings." World Journal of Diabetes **14**(3): 212–225.

American Diabetes, A. (2024). "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2024." Diabetes Care **47**(Supplement_1): S20–S42.

Association, A. D. (2018). "Standards of medical care in diabetes—2018 abridged for primary care providers." Clinical Diabetes.

AU - Ribeiro, M. T., et al. (2017). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Batista, G. E. A. P. A., et al. (2004). "A study of the behavior of several methods for balancing machine learning training data." ACM SIGKDD Explorations Newsletter **6**(1): 20–29.

Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5–32.

Breiman, L. (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1.

Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery.

Chawla, N. V., et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research **16**: 321–357.

Chen, Z., et al. (2022). "A Comprehensive Benchmark of Machine Learning Models for Diabetes Prediction." Journal of Biomedical Informatics **127**: 104005.

Cutler, A., et al. (2012). Random Forests. Ensemble Machine Learning: Methods and Applications. C. Zhang and Y. Ma, Springer**:** 157–175.

Federation, I. D. (2021). "IDF Diabetes Atlas 10th Edition."

Fernández, A., et al. (2018). "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary." Journal of Artificial Intelligence Research **61**: 863–905.

Fregoso-Aparicio, L., et al. (2021). "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review." Diabetology & Metabolic Syndrome **13**(1): 148.

Gupta, P. K. and R. K. Tripathi (2023). "Towards Automated Diagnostic Systems: A Review of AI in Healthcare." IEEE Reviews in Biomedical Engineering **16**: 123–139.

Hussein, A. S., et al. (2019). "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE." International Journal of Computational Intelligence Systems **12**(2): 1412–1422.

Ismail, L. and H. Materwala (2025). "IDMPF: intelligent diabetes mellitus prediction framework using machine learning." Applied Computing and Informatics **21**(1/2): 78–89.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." Science **349**(6245): 255–260.

Kangra, K. and J. Singh (2023). "Comparative analysis of predictive machine learning algorithms for diabetes mellitus." Bulletin of Electrical Engineering and Informatics **12**(3): 1728–1737.

Kavakiotis, I., et al. (2017). "Machine Learning and Data Mining Methods in Diabetes Research." Comput Struct Biotechnol J **15**: 104–116.

Lundberg, S. M. and S. I. Lee (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems.

Nissar, I., et al. (2024). "An Intelligent Healthcare System for Automated Diabetes Diagnosis and Prediction using Machine Learning." Procedia Computer Science **235**: 2476–2485.

Raschka, S. (2018). "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning." arXiv preprint arXiv:1811.12808.

Rashid, A. (2020). Diabetes Dataset, Mendeley.

Sahid, M. A., et al. (2024). "Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics." Plos one **19**(5): e0300785.

Sisodia, D. and D. S. Sisodia (2018). "Prediction of Diabetes using Classification Algorithms." Procedia Computer Science **132**: 1578–1585.

Waleed Khalil, K. B., Mohamed Mosadag (2025). "Early Detection of Chronic Kidney Disease (CKD)Using Machine Learning Algorithms." East Journal of Computer Science **1**(2): 1–9.

Wang, B. (2023). Employing Supervised Classification Algorithms for Diabetes Prediction, California State University San Marcos.

Yun, J. W. and S. M. Kim (2022). "Early Detection of Diabetes Mellitus for Preventing Complications." Journal of Preventive Medicine & Public Health **55**(5): 409–416.