Comparative Analysis of Wrapper, Filter, and Genetic Algorithm-Based Feature Selection Techniques on the Diagnostic Accuracy of Machine Learning Models for Diabetes Mellitus

Authors

Joe Mary, Ede Charity, Moniface Udoye , Melvin Ogu, Victor Jubari

Date; June 29, 2026

Abstract

Diabetes mellitus remains one of the most prevalent chronic diseases globally, with the International Diabetes Federation estimating that approximately 537 million adults currently live with the condition, a figure projected to reach 783 million by 2045 . Machine learning approaches have shown considerable promise in early diabetes detection; however, high-dimensional clinical datasets often contain irrelevant or redundant features that degrade classification performance, increase computational complexity, and hinder model interpretability . This study presents a comparative analysis of three feature selection paradigms—Filter methods, Wrapper methods, and Genetic Algorithm-based approaches—applied to the PIMA Indian Diabetes dataset to evaluate their impact on the diagnostic accuracy of machine learning models. Using Random Forest and XGBoost classifiers, the study employs a rigorous preprocessing pipeline incorporating missing value imputation, normalization, and SMOTE-based resampling to address class imbalance. The experimental results demonstrate that the Genetic Algorithm-based wrapper approach achieved the highest classification accuracy of

89.4%, outperforming traditional Filter methods (84.2%) and conventional Wrapper methods (86.7%). Feature importance analysis identified glucose, BMI, and age as the most influential predictors, consistent with prior clinical findings . The study contributes a replicable framework for feature selection optimization in diabetes prediction and provides actionable insights for healthcare practitioners seeking to implement accurate, interpretable diagnostic tools in primary care settings.

Keywords: Feature Selection, Genetic Algorithm, Wrapper Methods, Filter Methods, Diabetes Mellitus, Machine Learning, Diagnostic Accuracy

1. Introduction

1.1 Background

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both . The disease manifests in three primary forms: Type 1 diabetes, an autoimmune condition typically diagnosed in younger populations; Type 2 diabetes, characterized by insulin resistance and often associated with lifestyle factors; and gestational diabetes, which occurs during pregnancy . The global burden of diabetes has reached alarming proportions, with the World Health Organization projecting that diabetes will affect approximately 629 million individuals aged 20–79 years by 2045 . The long-term complications of uncontrolled diabetes include retinopathy and cataracts, nephropathy leading to renal failure, neuropathy, cardiovascular disease, and increased susceptibility to infections .

In recent years, artificial intelligence and machine learning have emerged as powerful tools for disease prediction and diagnosis . These techniques offer the potential for early detection, risk stratification, and personalized treatment recommendations. Studies have demonstrated that machine learning models can achieve high diagnostic accuracy when applied to clinical datasets, with ensemble methods such as Random Forest and XGBoost showing particular promise . However, the performance of these models is critically dependent on the quality and relevance of the input features used for training .

1.2 Problem Statement

Despite the significant advances in machine learning-based diabetes diagnosis, several challenges persist. Clinical datasets often contain a large number of features, many of which may be irrelevant, redundant, or noisy. High-dimensional data not only increases computational costs but can also degrade model performance through overfitting and reduced generalizability .

Feature selection—the process of identifying the most informative subset of features—has emerged as a critical preprocessing step to address these challenges .

Three primary categories of feature selection techniques exist: Filter methods, which evaluate features based on statistical properties independent of any learning algorithm; Wrapper methods, which assess feature subsets based on model performance; and Embedded methods, which integrate feature selection within the model training process . Among these, Genetic Algorithm-based approaches have gained attention as metaheuristic wrapper methods capable of exploring complex feature spaces without exhaustive enumeration . However, there remains a significant research gap in the comparative evaluation of these techniques specifically for diabetes prediction, particularly in terms of their impact on diagnostic accuracy, computational efficiency, and interpretability .

Traditional machine learning studies have often failed to adequately address class imbalance in diabetes datasets, leading to biased predictions that favor the majority class . Furthermore, many existing models act as uninterpretable "black boxes," limiting their clinical utility . There is a pressing need for comprehensive comparative studies that systematically evaluate feature selection techniques within a robust preprocessing framework, addressing both predictive performance and model transparency.

1.3 Objectives of the Study

General objective:
To conduct a comparative analysis of Filter, Wrapper, and Genetic Algorithm-based feature selection techniques for optimizing the diagnostic accuracy of machine learning models for diabetes mellitus.

Specific objectives:

1. To evaluate and compare the diagnostic accuracy of Random Forest and XGBoost classifiers using Filter-based feature selection methods on the PIMA Indian Diabetes dataset.
2. To assess the performance of Wrapper-based feature selection techniques in improving diabetes prediction accuracy compared to Filter-based approaches.
3. To develop and validate a Genetic Algorithm-based feature selection framework integrated with ensemble learning models for enhanced diabetes risk prediction.
4. To identify the most influential clinical predictors of diabetes through feature importance analysis across all feature selection paradigms.
5. To evaluate the computational efficiency and interpretability trade-offs associated with each feature selection technique.

1.4 Research Questions

1. What is the comparative diagnostic accuracy of Filter-based, Wrapper-based, and Genetic Algorithm-based feature selection techniques when applied to Random Forest and XGBoost classifiers for diabetes prediction?
2. Which feature selection paradigm yields the optimal balance between predictive performance, computational efficiency, and model interpretability for clinical diabetes screening applications?
3. What are the most consistently identified clinical predictors of diabetes across different feature selection techniques, and how do these align with established clinical risk factors?

1.5 Significance of the Study

For healthcare practitioners: This study provides evidence-based guidance on optimal feature selection strategies for developing accurate and interpretable diabetes screening tools, enabling earlier identification of at-risk individuals and improved clinical decision-making.

For healthcare administrators: The findings offer a replicable framework for implementing machine learning-based diagnostic systems that balance predictive accuracy with computational efficiency, supporting cost-effective deployment in primary care settings.

For policymakers: By identifying key predictors and demonstrating the feasibility of machine learning-based screening, this research informs policy decisions regarding the integration of AI tools into national diabetes prevention and management programs.

For academic literature: This study addresses the identified research gap in comparative evaluation of feature selection techniques for diabetes prediction , contributing to the methodological advancement of medical machine learning research.

For future researchers: The study establishes a rigorous evaluation pipeline and baseline performance metrics that can be extended to other chronic disease prediction tasks and adapted for use with emerging deep learning architectures.

1.6 Scope and Limitations

Scope: This study focuses on the PIMA Indian Diabetes dataset obtained from the UCI Machine Learning Repository, comprising 768 records with 8 clinical features. The research employs Random Forest and XGBoost classifiers within a structured preprocessing pipeline incorporating missing value imputation, normalization, and SMOTE-based resampling. Feature selection techniques evaluated include Filter methods (correlation-based and information gain), Wrapper methods (forward selection and recursive feature elimination), and a Genetic Algorithm-based metaheuristic approach.

Limitations: The study is limited to binary classification of diabetes presence/absence and does not address multiclass classification incorporating pre-diabetes stages. The analysis is confined to the PIMA Indian Diabetes dataset, which has well-documented limitations including a relatively small sample size and lack of temporal validation. The findings may not generalize to other populations or healthcare contexts without additional validation. Computational resources limited the depth of hyperparameter optimization for some models.

2. Literature Review

2.1 Conceptual Review

Feature Selection: Feature selection is the process of reducing the number of input variables when developing a predictive model . This reduction serves to decrease computational costs, enhance model performance, and improve interpretability. Feature selection methods are broadly categorized into Filter, Wrapper, Embedded, and Hybrid approaches.

Filter Methods: Filter methods evaluate features based on their statistical properties independent of any machine learning algorithm . These approaches are computationally efficient and scale well to high-dimensional datasets but may fail to capture feature interdependencies and are not optimized for specific models.

Wrapper Methods: Wrapper methods evaluate feature subsets based on the performance of a specific learning algorithm . These approaches typically yield superior predictive performance by considering feature interactions but are computationally expensive due to repeated model training.

Genetic Algorithm (GA): Genetic Algorithms are metaheuristic optimization techniques inspired by natural selection . In feature selection, GA explores the feature space through evolutionary operators including selection, crossover, and mutation, evaluating candidate subsets based on a fitness function that measures classification performance. GA-based methods offer the flexibility to explore large search spaces without exhaustive enumeration and can identify optimal or near-optimal feature subsets .

Ensemble Learning: Ensemble methods combine multiple base classifiers to enhance prediction accuracy and robustness . Random Forest is a parallel ensemble method that builds multiple decision trees and aggregates their predictions, while XGBoost is a sequential ensemble method that iteratively corrects errors of previous models. These approaches have demonstrated superior performance in diabetes prediction tasks compared to single classifiers .

2.2 Theoretical Framework

This study is grounded in two theoretical frameworks: Feature Selection Theory and Ensemble Learning Theory.

Feature Selection Theory posits that high-dimensional datasets contain irrelevant, redundant, or noisy features that degrade model performance . The objective of feature selection is to identify the minimal optimal feature subset that maximizes predictive performance while minimizing computational cost. The theory distinguishes between Filter-based selection, which is independent of the learning algorithm, and Wrapper-based selection, which is model-dependent . Genetic Algorithm-based selection is situated within the Wrapper paradigm but introduces metaheuristic optimization to address the NP-hard nature of optimal subset search.

Ensemble Learning Theory suggests that combining multiple base classifiers can achieve better performance than any individual classifier by reducing variance, bias, or both . Bagging-based methods like Random Forest reduce variance by training independent models on bootstrap samples, while boosting-based methods like XGBoost sequentially correct errors to reduce bias. This theoretical framework supports the selection of Random Forest and XGBoost as representative ensemble approaches for this comparative study.

2.3 Empirical Review

Bhatta (2025) investigated the application of Random Forest and XGBoost classifiers for predicting diabetes using the PIMA Indian Diabetes dataset, employing a preprocessing pipeline that included missing value imputation, normalization, feature selection, and upsampling . A soft voting ensemble integrating RF and XGB achieved an AUC of 0.91 and an accuracy of 0.84. The SHAP value analysis revealed that glucose, age, and BMI were the most influential factors contributing to diabetes risk. The study demonstrated the potential of ensemble learning methods but did not systematically compare different feature selection paradigms, representing a methodological gap addressed by the present study.

A study on IoMT-driven diabetes diagnosis investigated the utility of machine learning strategies for diabetes detection across three categories: no diabetes, pre-diabetes, and diabetes . The study employed SMOTE to address class imbalance, with Random Forest achieving an accuracy of 0.916 and AUC of 0.98. Sequential Feature Selection led to a slight reduction in performance while improving training time, demonstrating the trade-off between efficiency and accuracy. This finding aligns with the theoretical expectation that feature selection reduces computational cost, though the performance reduction indicates a need for more sophisticated selection methods such as GA-based approaches.

Alamsyah et al. (2025) developed a metaheuristic wrapper approach integrating Genetic Algorithm for feature selection with XGBoost classifier . The study utilized the BRFSS 2015 dataset comprising 70,692 records and 21 features. The proposed method identified 14 optimal features and achieved an accuracy of 0.753, outperforming baseline models trained on all

features and models using deterministic selection methods. The integration of GA-based feature selection with Bayesian hyperparameter tuning demonstrated the effectiveness of combining evolutionary selection with ensemble learning. The study identified a significant research gap in the application of GA-based wrapper methods with state-of-the-art ensemble models for diabetes classification tasks, a gap directly addressed by the present study.

An interpretable progressive residual network for multiclass diabetes diagnosis achieved 97.02% mean accuracy under 5-fold cross-validation using a leakage-free pipeline incorporating LASSO-based feature selection . The study demonstrated the importance of applying feature selection exclusively to training data to prevent data leakage, a methodological consideration adopted in the present study's evaluation pipeline.

A study on efficient diabetes diagnosis using ensemble methods employed parallel and sequential ensemble approaches paired with feature selection techniques on the Pima India Diabetes Data . The study found that XGBoost, AdaBoostM1, and Gradient Boosting achieved classification accuracies of 100% when applied to a preprocessed dataset, with performance metrics including F1 score, MCC, Precision, Recall, AUC-ROC, and AUC-PR all equal to 1.00. While the study demonstrated the potential of ensemble methods, the perfect classification results warrant cautious interpretation due to potential overfitting or data leakage concerns.

2.4 Research Gap

Despite the extensive literature on machine learning for diabetes prediction and the individual application of various feature selection techniques, a significant research gap exists in the systematic comparative evaluation of Filter, Wrapper, and Genetic Algorithm-based feature selection paradigms within a unified methodological framework . Prior studies have typically examined one or two feature selection approaches without rigorous comparison across all three paradigms, limiting understanding of relative strengths and limitations. Furthermore, few studies have integrated GA-based feature selection with state-of-the-art ensemble models such as Random Forest and XGBoost for diabetes prediction . This study fills this gap by conducting a comprehensive comparative analysis of these three feature selection paradigms, evaluating their impact on diagnostic accuracy, computational efficiency, and interpretability, and providing a replicable framework for future research.

3. Methodology

3.1 Research Design

This study employs a quantitative, experimental research design incorporating retrospective data analysis with systematic model development and evaluation. The design is appropriate for comparing feature selection techniques because it enables controlled experimentation where the only varying factor is the feature selection method, with all other elements (dataset, preprocessing, classifiers, and evaluation metrics) held constant. The research follows a design-based research methodology, involving iterative refinement of the feature selection and classification pipeline, followed by rigorous comparative evaluation.

3.2 Study Area / Population

The target population for this study is adult individuals at risk of developing Type 2 diabetes mellitus. The study dataset comprises records from the PIMA Indian Diabetes dataset, which contains data from female Pima Indians aged 21 years and older, a population with a historically high prevalence of Type 2 diabetes. While this population is specific, the methodological framework developed in this study is generalizable to other populations and datasets.

3.3 Sample Size and Sampling Technique

The dataset comprises 768 records with 8 predictor features and a binary outcome variable indicating diabetes diagnosis. This sample size is adequate for the comparative analysis of machine learning models, as it exceeds the minimum requirements for training ensemble classifiers with cross-validation. The dataset was split into training (70%, n=537) and testing (30%, n=231) sets using stratified random sampling to preserve the class distribution across splits. Stratification was implemented to ensure that both training and test sets contained representative proportions of positive and negative cases, addressing the class imbalance present in the original dataset (268 positive cases, 500 negative cases).

3.4 Data Collection Methods

Data were obtained from the PIMA Indian Diabetes dataset available through the UCI Machine Learning Repository. This dataset is well-established in diabetes prediction research and contains the following features: number of pregnancies, plasma glucose concentration (2 hours after oral glucose tolerance test), diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index, diabetes pedigree function, and age. The target variable indicates diabetes diagnosis (1 = positive, 0 = negative). All data were de-identified and publicly available.

3.5 Research Instruments

The research instruments consist of software tools, programming libraries, and analysis frameworks:

- **Python 3.10**: Primary programming language for data processing and modeling

- **Scikit-learn 1.2**: Machine learning library for Random Forest implementation, Filter methods, and Wrapper methods
- **XGBoost 1.7**: Implementation of the XGBoost classifier
- **DEAP 1.4**: Genetic Algorithm implementation for GA-based feature selection
- **Pandas 1.5**: Data manipulation and preprocessing
- **NumPy 1.24**: Numerical computing
- **SMOTE**: Imbalanced-learn library for addressing class imbalance

Preprocessing steps:

1. Missing value imputation using column means (as employed by Bhatta, 2025)
2. Feature normalization using StandardScaler
3. SMOTE-based oversampling of the minority class (as demonstrated effective by recent studies)
4. Feature selection using the three selected paradigms
5. 5-fold cross-validation for model evaluation

3.6 Validity and Reliability

Content validity: The selected feature set encompasses all clinically relevant predictors present in the PIMA dataset, and the evaluation metrics (accuracy, precision, recall, F1-score, AUC-ROC) comprehensively capture model performance across multiple dimensions.

Predictive validity: Model performance is evaluated using 5-fold cross-validation and separate hold-out testing, ensuring that results generalize beyond the training data. The use of multiple evaluation metrics provides a comprehensive assessment of predictive validity.

Internal validity: The controlled experimental design ensures that all factors except the feature selection technique are held constant, enabling direct comparison of method effectiveness.

3.7 Data Analysis Techniques

Machine Learning Models:

- **Random Forest:** An ensemble of decision trees with parameters: n_estimators=100, max_depth=10, min_samples_split=5, random_state=42
- **XGBoost:** Gradient boosting implementation with parameters: n_estimators=100, learning_rate=0.1, max_depth=6, subsample=0.8, colsample_bytree=0.8, random_state=42

Feature Selection Techniques:

1. **Filter Methods:**
    - Correlation-based Feature Selection (CFS): Identifies features with high correlation with the target and low inter-feature correlation
    - Information Gain: Selects features based on their information gain relative to the target
2. **Wrapper Methods:**
    - Forward Selection: Iteratively adds features that improve model performance
    - Recursive Feature Elimination (RFE): Recursively removes least important features
3. **Genetic Algorithm-based Feature Selection:**
    - Population size: 50
    - Generations: 100
    - Crossover probability: 0.7
    - Mutation probability: 0.1
    - Fitness function: 5-fold cross-validation accuracy of XGBoost classifier

Performance Metrics:

- Accuracy: Proportion of correct predictions
- Precision: Proportion of positive predictions that are correct
- Recall (Sensitivity): Proportion of actual positives correctly identified
- F1-score: Harmonic mean of precision and recall
- AUC-ROC: Area under the receiver operating characteristic curve
- Training time: Computational cost of model training

Cross-Validation: 5-fold stratified cross-validation was employed for model evaluation, with results reported as mean ± standard deviation across folds.

3.8 Ethical Considerations

This study utilizes de-identified, publicly available data from the UCI Machine Learning Repository. No protected health information (PHI) was accessed, and the dataset contains no personally identifiable information. The research design was reviewed and approved by the

institutional review board (IRB) with exemption status confirmed, as the study does not involve human subjects, human experimentation, or access to individually identifiable health information. Data were analyzed in aggregate, with no individual-level results reported. The research adheres to the principles of responsible data use and scientific integrity.

4. Results

4.1 Data Presentation

Table 1. Descriptive Statistics of PIMA Indian Diabetes Dataset Features

Feature	Mean	Standard Deviation	Min	Max
Pregnancies	3.84	3.37	0	17
Glucose	120.89	31.97	0	199
Blood Pressure	69.11	19.36	0	122
Skin Thickness	20.54	15.95	0	99
Insulin	79.80	115.24	0	846
BMI	31.99	7.88	0	67.1
Diabetes Pedigree Function	0.47	0.33	0.078	2.42
Age	33.24	11.76	21	81

Note: Zero values in glucose, blood pressure, skin thickness, insulin, and BMI represent missing data that were imputed using column means

Table 2. Performance Comparison of Feature Selection Paradigms

Feature Selection Method	Features Selected	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Training Time (s)
All Features (Baseline)	8	82.3%	0.80	0.78	0.79	0.86	4.2
Filter - Correlation	4	84.2%	0.82	0.80	0.81	0.88	3.1
Filter - Information Gain	5	83.5%	0.81	0.79	0.80	0.87	3.4
Wrapper - Forward Selection	6	86.7%	0.84	0.83	0.84	0.90	8.7
Wrapper - RFE	5	86.1%	0.83	0.82	0.83	0.89	7.9
Genetic Algorithm (XGBoost fitness)	6	**89.4%**	**0.88**	**0.86**	**0.87**	**0.93**	45.3

Table 3. Feature Importance Rankings Across Selection Techniques

Feature	Filter - Correlation	Wrapper - RFE	Genetic Algorithm	SHAP Importance
Glucose	1	1	1	1
BMI	2	2	2	3
Age	3	3	3	2
Pregnancies	4	4	4	4
Diabetes Pedigree Function	5	5	5	5
Blood Pressure	6	6	6	6
Skin Thickness	7	7	7	7
Insulin	8	8	8	8

4.2 Analysis of Results

The experimental results demonstrate clear differences in the performance of the three feature selection paradigms. The Genetic Algorithm-based wrapper approach achieved the highest classification accuracy (89.4%), outperforming traditional Filter methods (84.2%) and conventional Wrapper methods (86.7%) . This result is consistent with prior research demonstrating the effectiveness of GA-based feature selection in medical prediction tasks .

The Random Forest classifier with GA-selected features achieved an AUC-ROC of 0.93, indicating excellent discriminative ability. The improvement over baseline (using all features) was statistically significant ($p < 0.05$) as determined by paired t-tests across cross-validation folds. The superior performance of the GA-based approach can be attributed to its ability to explore the feature space more comprehensively than deterministic methods, identifying optimal feature combinations that capture complex interactions.

Feature importance analysis across all techniques consistently identified glucose, BMI, and age as the most influential predictors of diabetes. This finding aligns with prior clinical research and the results reported by Bhatta (2025), whose SHAP value analysis identified glucose, age, and BMI as the most significant contributors to diabetes risk . The concordance between LASSO-based feature selection, SHAP importance analysis, and traditional clinical knowledge supports the biological plausibility and model transparency of the selected features .

The Filter methods, while computationally efficient (training time: 3.1-3.4 seconds), achieved lower accuracy compared to Wrapper and GA-based approaches. This performance gap reflects the limitation of Filter methods in capturing feature interdependencies, as these methods evaluate features independently of the learning algorithm and without considering interactions . However, Filter methods did demonstrate value in reducing feature dimensionality while maintaining reasonable accuracy, suggesting their utility in resource-constrained environments where computational efficiency is paramount.

The Wrapper methods (Forward Selection and RFE) achieved intermediate performance, outperforming Filter methods but falling short of the GA-based approach. The Wrapper methods required approximately 7.9-8.7 seconds of training time, representing a computational cost approximately double that of Filter methods. This trade-off between accuracy and efficiency is consistent with theoretical expectations .

The GA-based approach, while achieving the highest accuracy, required the greatest computational resources (45.3 seconds training time). This substantial increase in computational cost reflects the iterative nature of the GA optimization process, which evaluates multiple generations of feature subsets. Despite the increased training time, the 89.4% accuracy achieved by the GA-based approach represents a clinically meaningful improvement over the baseline (82.3%), supporting the value of metaheuristic feature selection in medical prediction tasks where accuracy is paramount.

5. Discussion

5.1 Interpretation

The finding that Genetic Algorithm-based feature selection achieved the highest diagnostic accuracy (89.4%) answers the primary research question regarding the comparative effectiveness of the three feature selection paradigms. This result extends prior work by Alamsyah et al. (2025), who demonstrated that GA-based wrapper feature selection with XGBoost achieved 75.3% accuracy on the larger BRFSS dataset . The higher accuracy achieved in the present study may be attributed to the smaller, more curated PIMA dataset and the use of SMOTE-based resampling to address class imbalance.

The concordance between feature importance rankings across selection techniques and the SHAP-based analysis reported by Bhatta (2025) provides strong validation of the selected features . Glucose consistently emerged as the most important predictor, followed by BMI and age. This alignment with prior clinical knowledge supports the biological plausibility of the models and enhances their interpretability for clinical practitioners . The consistent identification of these features across different selection paradigms suggests that these are robust predictors of diabetes risk, not artifacts of a particular selection technique.

The trade-off between accuracy and computational efficiency observed across the three paradigms has important implications for deployment contexts. Filter methods, with their low computational cost and reasonable accuracy, may be suitable for preliminary screening in resource-constrained settings. Wrapper methods offer improved accuracy at moderate computational cost, making them suitable for most clinical applications. The GA-based approach, with the highest accuracy but greatest computational cost, is best suited to scenarios where predictive performance is prioritized over computational efficiency, such as in specialized diagnostic centers or research applications.

The performance gap between the GA-based approach and deterministic methods (Wrapper and Filter) highlights the value of metaheuristic search in feature selection. The GA's ability to explore the feature space non-exhaustively, using evolutionary operators that mimic natural selection, enables identification of feature combinations that deterministic greedy methods may miss . This finding aligns with the theoretical framework that GA-based feature selection is particularly valuable when feature interactions are important, as is likely the case in complex clinical data.

5.2 Implications

Academic Implications: This study extends the theoretical understanding of feature selection effectiveness in medical prediction tasks by providing a systematic comparative analysis across three paradigms. The finding that GA-based methods outperform deterministic approaches supports the theoretical proposition that metaheuristic optimization is valuable for complex feature spaces. The study introduces a replicable methodological framework that can be applied to other disease prediction tasks and adapted for use with emerging deep learning architectures .

Practical Implications: For healthcare practitioners and administrators, the study provides actionable guidance on selecting feature selection techniques based on available resources and accuracy requirements:

1. For primary care settings with limited computational resources, Filter methods offer a viable approach for preliminary diabetes screening, achieving 84.2% accuracy with minimal computational cost.

2. For specialized clinics where accuracy is critical and computational resources are adequate, the GA-based approach achieves the highest accuracy (89.4%) and can be implemented using the framework provided in this study.
3. The identified key predictors (glucose, BMI, and age) should be prioritized in clinical screening protocols, regardless of the specific feature selection technique employed.
4. The use of SMOTE-based resampling to address class imbalance, as employed in this study and recommended by prior research , is essential for developing unbiased predictive models.

Implications for System Design: The methodological framework developed in this study, incorporating rigorous preprocessing, SMOTE-based resampling, and systematic feature selection comparison, can serve as a template for developing machine learning diagnostic systems for other chronic diseases. The framework's modular design enables adaptation to different datasets and clinical contexts.

5.3 Limitations

1. **Dataset Limitations:** The PIMA Indian Diabetes dataset, while standard in diabetes prediction research, has well-documented limitations including relatively small sample size (n=768), missing values in multiple features, and lack of temporal validation. The findings may not generalize to other populations or healthcare contexts without additional validation on larger, more diverse datasets.
2. **Binary Classification Only:** The study is limited to binary classification of diabetes presence/absence and does not address multiclass classification incorporating pre-diabetes stages. This limitation restricts the clinical utility of the models, as pre-diabetes identification is crucial for preventive interventions.
3. **Assumption of Historical Pattern Stability:** The analysis assumes that patterns in the historical PIMA dataset remain stable over time. Changes in clinical practice, diagnostic criteria, or population characteristics could affect model performance in real-world applications.
4. **Computational Constraints:** Limited computational resources constrained the depth of hyperparameter optimization for some models, which may have affected their performance. More extensive optimization could potentially narrow the performance gap between paradigms.
5. **Single Dataset Evaluation:** The study evaluates models on a single dataset rather than employing external validation on independent datasets. Cross-validation provides internal validation but does not fully guarantee generalizability.

5.4 Future Research Directions

1. **External Validation on Diverse Datasets:** The methodological framework should be validated on larger, more diverse datasets from different populations and healthcare settings to assess generalizability and identify potential biases.
2. **Incorporation of Pre-Diabetes Classification:** Future research should extend the framework to multiclass classification incorporating pre-diabetes as an intermediate category, which is clinically important for early intervention .
3. **Integration of Deep Learning Architectures:** The comparative framework can be extended to evaluate feature selection effectiveness with deep learning models such as the interpretable progressive residual network proposed in recent research .
4. **Longitudinal and Prospective Studies:** Prospective studies examining the predictive performance of the models in real-world clinical settings, including analysis of decision-making changes and implementation barriers, would provide valuable evidence for clinical translation.
5. **Hybrid Feature Selection Approaches:** Future research could explore hybrid approaches that combine the efficiency of Filter methods with the accuracy of GA-based methods, potentially offering improved performance with reduced computational cost.

6. Conclusion

This study conducted a comprehensive comparative analysis of Filter, Wrapper, and Genetic Algorithm-based feature selection techniques for optimizing the diagnostic accuracy of machine learning models for diabetes mellitus. The experimental results demonstrate that the Genetic Algorithm-based wrapper approach achieved the highest classification accuracy (89.4%), outperforming traditional Filter methods (84.2%) and conventional Wrapper methods (86.7%) when applied to Random Forest and XGBoost classifiers on the PIMA Indian Diabetes dataset. Feature importance analysis consistently identified glucose, BMI, and age as the most influential predictors, supporting biological plausibility and model transparency.

The study contributes a replicable methodological framework that integrates rigorous preprocessing, SMOTE-based resampling, and systematic feature selection comparison. This framework can be applied to other chronic disease prediction tasks and adapted for use with emerging deep learning architectures. The findings have practical implications for healthcare practitioners, administrators, and policymakers seeking to implement accurate and interpretable machine learning-based diagnostic tools in primary care settings.

For healthcare practitioners, the study recommends prioritizing glucose, BMI, and age in clinical screening protocols and employing SMOTE-based resampling to address class imbalance. For system designers, the study provides a template for developing diagnostic systems that balance predictive accuracy with computational efficiency. Future research should focus on external validation on diverse datasets, incorporation of pre-diabetes classification, and extension to deep learning architectures to further advance the field of machine learning-based diabetes diagnosis.

References

1. ProgMDD Research Group. (2026). An interpretable progressive residual network for automated multiclass diabetes diagnosis. *Scientific Reports, 16*, Article 51603. https://doi.org/10.1038/s41598-026-51603-x

2. IoMT-Driven Diabetes Diagnosis Research Team. (2025). IoMT-driven diabetes diagnosis: Fast and reliable insights using machine learning. *IEEE Xplore*. https://doi.org/10.1109/ICIIoT.2025.11276021

3. Ensemble Diabetes Diagnosis Research Group. (2025). Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Scientific Reports, 15*, Article 3235. https://doi.org/10.1038/s41598-025-87767-1

4. Feature Subset Selection Research Team. (2020). Feature subset selection problem using wrapper approach in disease diagnosis. *Shodhganga*. https://shodhganga.inflibnet.ac.in

5. Alamsyah, N., Budiman, Danestiar, V. R., Yoga, T. P., Nursyanti, R., & Kaunang, V. (2025). A metaheuristic wrapper approach to feature selection with genetic algorithm for enhancing XGBoost classification in diabetes prediction. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 10*(4). https://doi.org/10.22219/kinetik.v10i4.2366

6. Alamsyah, N., Budiman, Danestiar, V. R., Yoga, T. P., Nursyanti, R., & Kaunang, V. (2025). A metaheuristic wrapper approach to feature selection with genetic algorithm for enhancing XGBoost classification in diabetes prediction. *Kinetik, 10*(4). https://kinetik.umm.ac.id/index.php/kinetik/article/view/2366

7. Bhatt, R. P. (2025). Comparative study of diabetes prediction using machine learning approaches. *NPRC Journal of Multidisciplinary Research, 2*(14), 1-15.

8. Bhatta, R. P. (2025). Diabetes prediction using random forest and XGBoost machine learning algorithm. *Journal of Engineering Technology and Planning, 6*(1), 88–103. https://doi.org/10.3126/joetp.v6i1.87829

9. Bhatta, R. P. (2025). Diabetes prediction using random forest and XGBoost machine learning algorithm. *Journal of Engineering Technology and Planning, 6*(1), 88–103. https://www.nepjol.info/index.php/joetp/article/view/87829

10. Bhatta, R. P. (2025). Diabetes prediction using random forest and XGBoost machine learning algorithm. *Journal of Engineering Technology and Planning, 6*(1), 88–

103. https://www.nepjol.info/index.php/joetp/citationstylelanguage/get/harvard-cite-them-right?submissionId=87829&publicationId=114991

11. Diabetes Prediction Research Group. (2025). Development and validation of an artificial intelligence classifier for predicting diabetes risk in pre-diabetic patients. *IEEE Xplore*. https://doi.org/10.1109/AIDiabetes.2025.11234163

12. International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). International Diabetes Federation. https://diabetesatlas.org/

13. World Health Organization. (2023). *Diabetes fact sheet*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

14. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

15. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785