Research Article

A Hybrid Recursive Feature Elimination Principal Component Analysis Framework With Optimized Support Vector Machine For Accurate Diabetes Prediction

Rahul Ranjan[1], Mohd Haroon[2]

[1]Department of CSE Integral University, Lucknow, India. Email: rrtiwari88@gmail.com
[2]Department of CSE Integral University, Lucknow, India. Email: mharoon@iul.ac.in

Abstract

Timely diagnosis of diabetes supports earlier clinical action, yet standard machine-learning pipelines often struggle with overlapping features, inter-correlated clinical measurements, and models whose decisions are hard to explain. In this paper, we present an explainable hybrid framework based on Recursive feature elimination – Principal Component Analysis – Radial Basis Function Support Vector Machine (RFE-PCA-RBF-SVM). The study exploits Pima Indians Diabetes Dataset (PIDD), which consists of 768 cases and eight clinical attributes. Our framework comprises handling of invalid data based on median imputation, Min-Max normalization of the dataset, RFE feature selection, PCA dimensionality reduction, and hyperparameter optimization of RBF-SVM. The experiment was designed to use 80:20 stratified train-test scheme as well as 10-fold stratified cross-validation, ablation analysis, and statistical validation methods. It was found out that hybrid framework performed better than the base model in terms of measured characteristics and achieved accuracy of 71.08% (standard deviation = 2.71%). Considering that Histogram Gradient Boosting surpassed the other methods analyzed in terms of the predictive performance of the model, the hybrid framework proposed in this study managed to remain highly effective offering a well-structured process for feature optimization and nonlinear classification. In this context, the explainability based on the SHAP methodology determines the importance of Glucose, BMI, Age, and Diabetes Pedigree Function for the model prediction results. Consequently, the proposed method provides a neat, statistically verified, and interpretable technique for diabetes prediction, demonstrating the significance of the combination of predictive evaluation with component-wise ablation and explainability analysis.

Keywords: Diabetes Prediction; Pima Indians Diabetes Dataset; Recursive Feature Elimination; Principal Component Analysis; RBF-SVM; Hybrid Machine Learning; Feature Optimization; Explainable Artificial Intelligence; SHAP; Statistical Validation; Cross-Validation.

1.0 Introduction:

1.1 Background and Research Context

Diabetes mellitus is a long-term metabolic condition that places heavy demands on healthcare systems worldwide because of how common it is, how many complications it produces, and the resulting cost of care. When blood glucose stays poorly controlled over time, it can progressively damage the heart, kidneys, nerves, and eyes, and raise the risk of stroke[1]. The consequences of the late diagnosis are extremely important because many patients remain undetected as far as the disease shows very fuzzy symptoms during the first stages of its progress. As a result, effective risk prediction becomes an important component of preventive medicine and helps at-risk people to undergo further diagnostic process and treatment. This research explains the notion of risk assessment and the development of machine learning supported systems for risk assessment.Magic of AI and machine learning results in numerous opportunities for support of data-based medical decisions. Unlike conventional analytical methods based on logical guessing of the relations between clinical parameters and outcomes, properly tuned supervised ML algorithms may uncover complex relationships between clinical parameters and outcomes based on the historical data.A wide range of classifiers has been tested on this problem, including Logistic Regression[2], tree-based models such as Decision Trees and Random Forest, probabilistic learners like Naïve Bayes, distance-based methods such as KNN, margin-based Support Vector Machines, and neural-network approaches.In fact, SVM is often determined to be a suitable tool for binary classification in medicine due to its maximum-margin method of learning and its ability to predict nonlinear functions through kernel functions that help SVM in solving its challenges.However, it can be said that the performance of an SVM classifier entirely depends on the type and quality of its features, so it would be incorrect to state that it is enough to select a powerful classifier[3].

1.2 Challenges in Machine-Learning-Based Diabetes Prediction

While modern technologies like machine learning have evolved, achieving full automation of diabetes prediction remains problematic because clinical datasets are often compiled of incomplete entries, excessive

features, correlated predictors, various measurement scales, and class imbalance, all of which obstruct model learning and produce unreliable or biased decision contours[4]. One illustrative example of this issue is Pima Indians diabetes dataset that demonstrates all of these problems perfectly since it provides a sample of 768 entries of patients suffering from eight clinical characteristics of patients afflicted by diabetes and a binary dependent variable including 500 records with individuals without diabetes and 268 instances concerning patients with diabetes.Another issue is that some clinical indicators have physiologically unrealistic values of zero. This suggests that different zero values in the database are not synonymous with any physiological facts and need to be processed before using the data for training[5]. In addition, one has to pay attention to the fact that pediatrics has its own system of measurement, thus, normalization is essential when applying distance-and kernel-based models. Aside from that, correlation analysis of clinical features needs the application of dimensionality expansion methods to obtain simpler and more stable representations.This problem is therefore much wider than choosing a classifier. The task of diagnosing diabetes has to do with building an integrated learning pipeline that takes into account the data quality, important features, dimensionality, properties of classifiers, and validation methodology[6].

1.3 Limitations of Existing Machine Learning Approaches

Various research works have been conducted for the purpose of comparison of different algorithms used in diabetes classification. The studies are informative because they provide various proofs concerning the efficiency of algorithms but simply comparing the algorithms does not clarify the reason for efficient functioning of each of the ones mentioned above[7].

One of the main drawbacks is that the entire set of features is used directly, which means that presence of some irrelevant and not informative features makes it possible for the learning process to have a lot of unnecessary variability. At the same time, usage of correlated features produces duplicity, which means that quality of results obtained gets worse.

Another limitation concerns the issue of dimensionality reduction. Even though PCA and some other methods of dimensionality reduction are widely used individually, application of such methods without previous checking of what features are relevant is inappropriate because they induce variance with the help of features that might not result in improvement of the task in question. Moreover, feature selection will eliminate irrelevant features but keep the correlated ones[8].

1.4 Motivation for Recursive Feature Elimination Feature selection means leaving out unnecessaryinformation but keeping those variables that are essential for predicting a specific outcome. In terms of wrapping approaches, Recursive Feature Elimination (RFE) is the most relevant of these techniques because it provides an evaluation of feature importance through the elimination of variables that are least significant in the very process of operation until a smaller, nevertheless satisfactory number of features is left as a result.The solution stated in this paper suggests using RFE before applying PCA for a purpose. RFE does not apply PCA for every feature raw. Instead, it aims at selecting only the features that are relevant for prediction before applying PCA to them.This step indicates that the research introduces an important methodological question which states whether it is possible to get better results if feature selection and PCA are applied together instead of using only one of these methods.According to the paper, the combination of RFE and PCA is still not studied thoroughly in diabetes prediction[9].

1.5 Motivation for Principal Component Analysis

While RFE trims the feature set, it does not by itself remove correlation among the variables that remain. PCA addresses this gap by re-expressing the retained variables as a smaller set of uncorrelated components, simplifying the feature space without discarding meaningful variance.The technique has around 95% of total variance, which means there is a balance between the need for reducing the number of variables and keeping their information. The main goal of this approach is to deal with multicollinearity issues that stabilize calculations and improve quality of features presented.

Thus, PCA will not be used just to reduce the amount of variables in the project but is meant to be the second step in feature optimization after the application of RFE method.

1.6 Research Gap

1. The thorough review of existing literature results in identifying the following gap in research: the majority of the research on diabetes prediction focuses on classifier performance thus little attention is paid to the real contribution of the feature representation.
2. Moreover, it is known that both PCA and RFE are effective separately, but not enough focus was put on implementation of both methods together.
3. In addition, while reports claim that improvements in performance were achieved, it is complicated to determine what leads this improvement. Thus, it is necessary to conduct an ablation study in order to compare different systems.

4. Finally, in addition to analyzing the performance of a separate model, health prediction allows for concluding about the overall significance of the entire predictive system.
5. Furthermore, it is important for the predictive system to be interpretable, which can be achieved through utilization of SHAP method that allows for evaluating the significance of each individual predictor.

1.7 Main Contributions

The main contributions of this study are as follows:
(1) The process of hybrid feature optimization consists of a stepwise hybrid RFE-PCA approach where wrapper based feature selection is performed followed by orthogonal dimensionality reduction.
(2) An RBF-SVM classifier is deployed for the task of learning the non-linear classification boundaries based on the optimized feature representation.
(3) An explicit ablation study has been performed to establish the baseline, RFE-only, PCA-only, and hybrid RFE-PCA configurations, enabling the quantification of each contribution without making the assumption.
(4) An exhaustive evaluation of the performance of the system with respect to Accuracy, Precision, Recall, F1-Measure, MCC, Area Under the curve, Confusion Matrices, computational execution time, and cross-validation is performed.
(5) The SHAP-based approach has been adopted to improve the interpretability of the predictions made by the model.
(6) The study contains a detailed description of the software environment, random seed, dataset details, train-test split method, stratified cross-validation, and optimization method.

2. Related Work and Research Gap Analysis

Because computational models can examine several clinical variables at once and surface the complex relationships behind a diagnosis, machine learning has increasingly become a central tool in diabetes-prediction research. The amount of literature dedicated to the issue is quite large, as different types of models have been applied, such as linear classifiers, probabilistic models, instance-based learning, decision trees, ensemble methods, neural networks, and kernel-based classifiers. Nevertheless, the studies emphasize that the quality of the input features is equally, if not more, important for the success of the model as compared to the choice of the classifier[3].In this line, the current paper has been inspired by the discovery of the fact that many relevant methodological issues are usually studied separately. Thus, various techniques are applied for feature extraction, dimensionality reduction, and non-linear classification, as well as for the interpretation of the results with the use of explainable AI technologies. Nevertheless, it is still unclear whether there is a universal approach that would integrate those parts and evaluate their singular and combined effectiveness through the ablation procedure.

2.1 Machine Learning Approaches for Diabetes Prediction

In diabetes prediction, machine learning methods have progressed from using common statistical classifiers to now employing techniques such as hybrid systems, deep learning, ensemble learning, and interpretable predictive models. Conventional methods such as K-nearest neighbors, logistic regression, and naïve bayes are upheld as standard classifiers because of their simplicity, efficiency, and high rate of interpretation. Nonetheless, their performance could be limited in cases where there are nonlinear relationships among the clinical variables, especially in cases with complicated data interaction patterns[4].
Tree-based and ensemble techniques like random forest and gradient boosting have achieved good performance because they are adept at modeling nonlinear interactions and dependencies between variables. Random forest uses an ensemble of different types of decision trees which can also analyze feature importance. Gradient boosting can also apply improvements in each weak learner making its classification results even better.
Nonetheless, the outcome produced by the models differs significantly when it depends on feature representation, validation approach, preprocessing of the data, class distribution, and hyperparameters choice. Thus, comparison between different methods without analyzing the feature representation would not lead to a full understanding of how the models behave.

2.2 Feature Selection for Diabetes Prediction

The procedure of selecting appropriate features is crucial in the analysis of data related to health care as clinical datasets are likely to contain features with different predictive power. The presence of weakly informative or duplicated features can rise the complexity of the model as well as deteriorate generalization[10].

Modern feature selection techniques can be divided into three main groups:

1. Filter methods
2. Wrapper methods
3. Embedded methods

Filter methods detect significant features based on statistical criteria such as correlation, mutual information, or variance. A drawback of such approaches is their limited capacity to directly optimize the features' subsets with respect to a particular classifier.Wrapper techniques are employed in order to analyze subsets of features according to their predictive capability. Despite being much more complex from the computational point of view, these methods allow obtaining closer correlation between feature selection and the target learning algorithm.Embedded methods imply that feature selection is incorporated into the learning process. This can be seen in regularized regression and tree-structured feature importance techniques.Recursive Feature Elimination can be attributed to wrapper techniques. The essence of RFE is that features are removed iteratively according to their importance until getting the desired number of features [11].

2.3 Dimensionality Reduction and PCA-Based Learning

The aim of dimensionality reduction is to develop a compact form of the whole feature space. The most popular of the methods is Principal Component Analysis as the method transforms correlated variables into orthogonal ones.

For a particular feature matrix X, PCA generates a new representation:

Z = XWk

where the eigenvectors are included in Wk.

The total explained variance can be shown as follows:

$$V(k) = \sum_{i=1}^{k} \lambda i \; / \sum_{i=1}^{d} \lambda i$$

where λi refers to the eigenvalue of the ith principal component.

According to the method presented in the research, PCA is adapted in a way to keep about 95% of the total explained variance as per the experimental design given in the research paper.

However, it should be mentioned that PCA does not find out whether the initial variables are good for predicting the target class. As PCA maximizes variance and not classification relatedness, one should be careful, as while PCA can decrease gene space and reduce correlation, applying PCA to the initial gene space will preserve good and high variation that may be not the best one for the prediction outcome[12].

Therefore:

$$RFE \rightarrow Predictive\ Feature\ Relevance$$

while:

$$PCA \rightarrow Orthogonal\ and\ Compact\ Representation$$

The combination is designed to exploit the complementary strengths of both techniques.

2.4 Hybrid Feature Engineering for Predictive Healthcare

Current machine learning studies place increased importance on hybrid feature engineering methods instead of focusing on one method of pre-processing or transformation only. Hybrid methods are those that try to provide different functions in a complementary manner such as:

1. Feature identification;
2. Dimensionality exclusion;
3. Feature extraction;
4. Normalization;
5. Hyperparameter optimization;
6. Non-linear classification.

This, however does not mean that a combination of several methods makes a scientific contribution. The hybrid method should show that every component has a certain value contribution to the result achieved[13].

The existing disadvantage of the hybrid methods studied in machine learning is often that they fail to assess the results of the components used in the hybrid methods. When a hybrid method shows better results than a certain method a baseline, it is often not clear which component is responsible for that improvement: selection of the features, dimensionality reduction method, the classifier used, hyperparameter optimization, or a combination of several ideas.

To address this issue, the proposed study uses an ablation design involving four configurations:

$$BaselineBaseline + RFEBaseline + PCABaseline + RFE + PCA$$

This design enables the contribution of each component to be evaluated independently and collectively.

The scientific value of the proposed framework therefore lies not only in integrating RFE and PCA but also in experimentally validating the necessity of their sequential combination.

2.5 Kernel-Based Learning for Diabetes Prediction

Support Vector Machines learn a decision boundary by maximizing the margin that separates the two outcome classes. When the classes cannot be separated linearly, a kernel function implicitly projects the data into a higher-dimensional space where separation becomes possible[14].

The proposed framework employs the Radial Basis Function kernel:

$$(xi, xj) = exp(-\gamma|xi - xj|\hat{}2)$$

where γ controls the influence of individual training observations.

The optimization objective of the soft-margin SVM can be expressed as:

$$w, b, \xi min21|w|2 + Ci = 1\sum n\xi i$$

Subject to:

$$y(wT\phi(xi) + b) \geq 1 - \xi i$$

WhereC controls the trade-off between margin maximization and classification error.

The usefulness of SVM for healthcare classification depends significantly on the input representation. A poorly structured or redundant feature space may reduce the effectiveness of nonlinear kernel learning. Consequently, the proposed work investigates whether an RFE–PCA optimized representation can provide a more suitable input space for RBF-SVM classification.

This is a critical aspect of the scientific contribution: the proposed SVM is not considered independently, but as the final learning component of a sequential feature optimization and nonlinear classification framework.

2.6 Explainable Artificial Intelligence in Diabetes Prediction

While predictive performance remains important, the use of machine learning in healthcare also raises questions concerning transparency and interpretability. A high-performing classifier may be difficult to trust if the factors contributing to individual predictions remain unclear [34].

Explainable AI techniques try to close this gap by exposing how a model arrives at its predictions. Researchers have increasingly turned to tools such as LIME, SHAP, permutation-based importance scores, partial dependence plots, and counterfactual reasoning to achieve this.

Among these approaches, SHAP provides a theoretically grounded method for quantifying feature contributions to individual predictions. The SHAP formulation can be represented as:

$$(x) = \phi 0 + j = 1\sum M\phi j$$

where:

1. ϕ0 represents the baseline prediction;
2. ϕj represents the contribution of feature j;
3. M denotes the number of explanatory features.

SHAP can provide both:

1. Global explanations, identifying variables that contribute most strongly across the dataset.
2. Local explanations, showing why a particular observation receives a particular prediction.

However, explainability should be based on the actual trained model and observed data, rather than manually assigned feature-importance values. Accordingly, the proposed framework integrates SHAP as a post-hoc explanation stage for interpreting the predictions generated by the final trained model.

2.7 Exhaustive Literature Review and Comparative Analysis

Table 1. Comparative Review of Representative Studies on Machine-Learning-Based Diabetes Prediction

Study	Data / Application	Feature Engineering	Primary Model	Validation / Evaluation	Explainability	Major Findings and Limitations
Logistic Regression-based approaches	Clinical diabetes data	Standard preprocessing	LR	Accuracy, Precision, Recall	Limited	Interpretable baseline but limited nonlinear modelling capability
KNN-based approaches	Diabetes benchmark data	Normalization	KNN	Accuracy and F1-score	No	Simple and effective for local similarity; sensitive to feature scaling and neighborhood selection
Naïve Bayes-based approaches	Clinical datasets	Basic preprocessing	NB	Accuracy, Recall	Limited	Computationally efficient but relies on conditional independence assumptions

Random Forest-based studies	Diabetes prediction	Feature importance / preprocessing	RF	Accuracy, F1, ROC-AUC	Partial	Strong nonlinear modelling but can require substantial ensemble complexity
Gradient Boosting approaches	Clinical prediction	Scaling and tuning	GB/HGB/XGBoost	Accuracy, ROC-AUC	Limited/ Partial	Strong predictive performance but interpretation may remain challenging
ANN-based studies	Diabetes and healthcare data	Normalization	ANN	Accuracy, F1-score	Limited	Learns nonlinear relationships but may require careful tuning and larger datasets
Deep learning approaches	Healthcare datasets	Automated feature learning	CNN/LSTM/Hybrid	Accuracy and related metrics	Emerging	Powerful representation learning but increased computational and interpretability challenges
SVM-based studies	PIDD and clinical datasets	Scaling / conventional preprocessing	Linear/ RBF-SVM	Accuracy, Precision, Recall	Limited	Effective for nonlinear classification but highly dependent on feature representation and parameter selection
Feature-selection-based studies	Diabetes prediction	Filter/wrapper/embedded selection	Various ML models	Accuracy, F1, ROC-AUC	Rare	Improves feature relevance but may not address correlation among retained variables
PCA-based studies	Clinical classification	PCA	Multiple classifiers	Accuracy, dimensionality	Rare	Reduces redundancy but maximizes variance rather than direct class relevance
Hybrid ML studies	Diabetes prediction	Multiple preprocessing stages	Ensemble/hybrid models	Multiple metrics	Limited	Hybridization may improve performance, but component-wise validation is often insufficient
XAI-oriented healthcare studies	Medical classification	Model-dependent or model-agnostic	Various ML/DL models	Predictive + explanatory analysis	SHAP/LIME	Improves transparency but may not jointly optimize feature selection and dimensionality
Proposed Study	Pima Indians Diabetes Dataset	RFE + PCA	Optimized RBF-SVM	Accuracy, Precision, Recall, F1, MCC, ROC-AUC, CV, ablation, statistical analysis	SHAP	Sequential feature optimization with component-wise ablation and explainable nonlinear classification

2.8 Comparative Synthesis of Existing Approaches

Table 2. Methodological Comparison and Positioning of the Proposed Framework

Methodological Dimension	Conventional ML Studies	Feature Selection Studies	PCA-Based Studies	Kernel-Based Studies	XAI-Based Studies	Proposed Framework
Data Preprocessing	YES	YES	YES	YES	YES	YES

Explicit Feature Selection	Limited	YES	Limited	Limited	Variable	YES
Dimensionality Reduction	Rare	Rare	YES	Limited	Limited	YES
Sequential RFE → PCA	Rare	Rare	Rare	Rare	Rare	YES
Nonlinear Kernel Learning	Limited	Variable	Variable	YES	Variable	YES
Hyperparameter Optimization	Variable	Variable	Variable	Variable	Variable	YES
Ablation Study	Rare	Limited	Limited	Rare	Rare	YES
Cross-Validation	Variable	Variable	Variable	Variable	Variable	YES
Statistical Validation	Limited	Limited	Limited	Limited	Limited	YES
SHAP-Based Interpretation	Rare	Rare	Rare	Rare	YES	YES
Computational Complexity Analysis	Rare	Rare	Rare	Limited	Rare	YES

The purpose of this comparison is not to claim that no previous study has combined any of these methods [35]. Rather, the research contribution should be carefully stated as the systematic integration and validation of these components within the experimental framework of this study**.**

3. Proposed Methodology:

3.1 Overview of the Proposed Framework

This study proposes an explainable hybrid machine learning framework for diabetes prediction that integrates data preprocessing, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and an optimized Radial Basis Function Support Vector Machine (RBF-SVM)**.** The central objective is to construct a compact and discriminative feature representation before nonlinear classification while maintaining the interpretability of the resulting predictive model[15,16,17].

Let the original clinical dataset be represented as:

$D = \{(xi, yi)\} = 1\ to\ N$..... (1)

where:

$xi = [xi1, xi2, \dots, xid]T \in Rd$ .....(2)

denotes the feature vector of the ith patient, and$yi \in \{0,1\}$ represents the corresponding class label, where 0 denotes the non-diabetic class and 1 denotes the diabetic class.For the Pima Indians Diabetes Dataset:N=768,d=8[18,19].

The proposed framework learns a predictive mapping:

$f: Rd \rightarrow \{0,1\}$ ..... (3)

such that:$y\hat{}i = f(xi)$..... (4)

The complete methodology can be represented as:

where: D→Dp→Dr → $Dz \rightarrow Y \rightarrow \emptyset$

1. P represents data preprocessing;
2. R represents RFE-based feature selection;
3. T represents PCA transformation;
4. S represents optimized RBF-SVM classification;
5. E represents explainability analysis;
6. Φ denotes the explanatory feature contributions.

3.2 Dataset Representation and Problem Formulation

The dataset is represented by the feature matrix:

X∈RN×d and target vector: y∈{0,1}NThus:D=(X,y) ..... (5)

The feature matrix can be expressed as:$\begin{matrix} x11 & x12 & x1d \\ x21 & x22 & x2d \\ xn1 & xn2 & xnd \end{matrix}$.....(6)

The objective is to learn an optimized predictive function:

$f* = argf \in Fmin(y, f(X))$ .....(7)

where L represents a classification loss function and F denotes the hypothesis space.

However, instead of directly applying a classifier to X, the proposed framework first constructs an optimized representation[20,21]:

$X* = (R(P(X)))$ .....(8)

Therefore, the final predictive function becomes:

$y\hat{} = fSV(X*)$ .....(9)

This representation-learning strategy forms the core methodological contribution of the study.

3.2 Data Preprocessing

Data preprocessing is performed before feature selection and model training to improve data quality and ensure that all features are represented on a comparable scale.

The preprocessing pipeline consists of:

1. Identification of physiologically invalid values;
2. Missing-value treatment;
3. Data normalization;
4. Training and testing data separation.

$The\ preprocessing\ operation\ is\ represented\ as: Xp = (X)$.....(10)

3.3. Missing and Invalid Value Treatment

Let xij represent the jth feature of the ith observation. For selected clinical variables, zero values may represent invalid or unavailable physiological measurements.

An indicator function can be defined as[22,23,24]:

$mij = \{1,0, if\ xij\ is\ invalidotherwise$ .....(11)

For invalid observations, the replacement value can be estimated using the training-set mean:

$x{\sim}j = nj1i = 1\sum njxij$.....(12)

where only valid observations are used to calculate the estimate.

The corrected value becomes:

$xij* = \{x{\sim}j, xij, mij = 1mij = 0$ .....(13)

To avoid data leakage, the imputation statistics must be calculated exclusively from the training partition and subsequently applied to validation and test samples[25,6].

3.4 Feature Normalization

Because the clinical variables are represented using different numerical ranges, Min-Max normalization is applied[19]:

$xij' = xjmax - xjminxij - xjmin$.....(14)

where:

1. $xjmin\ is\ the\ minimum\ training\ value\ of\ feature\ j;$
2. $xjmax\ is\ the\ maximum\ training\ value\ of\ feature\ j.$

After transformation:

$xij' \in [0,1]$.....(15)

The normalized matrix is:

$Xn = [xij']$..... (16)

The normalization parameters are fitted only on training data:

$(xjmin, xjmax) = Fi(Xtrain)$ ..... (17)

and then applied consistently to validation and testing data.

3.5 Recursive Feature Elimination

The first feature optimization stage is Recursive Feature Elimination.

Let the complete feature set be:

$(0) = \{f1, f2, \dots, fd\}$ ..... (18)

At iteration t, a learning estimator assigns an importance value:wj(t) to every remaining feature fj.The least informative feature is determined as:

$fmi(t) = argfj \in F(t)minwj(t)$..... (19)

The feature subset is then updated:

$$(t+1) = F(t) \setminus \{fmin(t)\} \text{..... (20)}$$

This recursive procedure continues until k features remain:

$|(T)| = k$ .....(21)

$$The\ RFE\ optimization\ objective\ can\ be\ expressed\ as: F* = argF' \subseteq FmaxQ(F')$$
$$subject\ to: |F'| = k\ where\ Q\ denotes\ a\ predictive\ quality\ criterion.$$

$$The\ reduced\ feature\ matrix\ is\ represented\ as: XRFE = XnS$$
$$where: S\ \{0,1\}d \times k\ \ is\ a\ feature - selection\ matrix. The\ role\ of\ RFE\ is\ therefore:$$
$$Xn \rightarrow XRFE\ \ which\ reduces\ the\ original\ feature\ representation$$
$$while\ retaining\ variables\ with\ higher\ predictive\ relevance.$$

3.5 Principal Component Analysis

Although RFE reduces the feature space, correlations may remain among the selected variables. Therefore, PCA is subsequently applied to construct an orthogonal representation[27,28,29].

$Let: XRFE \in RN \times k\ The\ centered\ feature\ matrix\ is: Xc = XRFE -$
$1\mu T\ where\ \mu\ denotes\ the\ feature\ mean\ vector$ .....(22)

$The\ covariance\ matrix\ is: C = N - 11XcTXc\ The\ eigenvalue\ decomposition\ is: Cvj = \lambda jvj\ where:$

1. λj denotes the jth eigenvalue;
2. vj denotes the corresponding eigenvector.

The eigenvalues are ordered such that:λ1≥λ2≥⋯≥λk

$$The\ cumulative\ explained\ variance\ for\ q\ principal\ components\ is: CEV(q) = \sum_{j=1}^{k} \lambda j / \sum_{j=1}^{q} q\ \lambda$$
$$The\ smallest\ number\ of\ components\ satisfying: CE(q) \geq 0.95\ is\ selected. Let: Vq = [v1, v2, \dots, vq]$$
$$The\ PCA - transformed\ representation\ is: Z = XcVq\ Thus: XRFE$$
$$\rightarrow Z\ The\ resulting\ feature\ space\ satisfies: q \leq k$$
$$\leq d\ and\ provides\ an\ orthogonal\ representation\ for\ subsequent\ classification.$$

3.6 Methodological Strength of the Proposed Framework

The scientific strength of the proposed methodology lies in the interaction of its individual stages[30,31,32]:

$Data\ QualityPreprocessing \rightarrow Feature\ RelevanceRFE \rightarrow Compact\ RepresentationPCA \rightarrow$
$Nonlinear\ LearningRBF - SVM \rightarrow Parameter\ OptimizationGrid\ Search \rightarrow InterpretabilitySHAP$

The corresponding methodological contribution can be expressed as:

Robust Prediction=f(Data Quality,Feature Relevance,Redundancy Reduction,Nonlinear Learning,Parameter Opt imization,Explainability) Importantly, the proposed study does not assume that the combination of these stages automatically improves performance. Their contribution is subsequently examined through the experimental ablation framework[33,34]:

$$M0: Baseline\ M1: RFE\ M2: PCA\ M3: RFE + PCA$$
$$The\ relative\ contribution\ of\ each\ component\ can\ then\ be\ measured\ as:$$
$$\Delta RFE = (M1) - M(M0)\ \Delta PCA = M(M2) - M(M0)$$
$$and:$$
$$\Delta Hybrid = (M3) - M(M0)$$
$$where\ (\cdot)\ represents\ a\ selected\ performance\ metric\ such\ as\ Accuracy, F1 - score, MCC, or\ ROC - AU$$

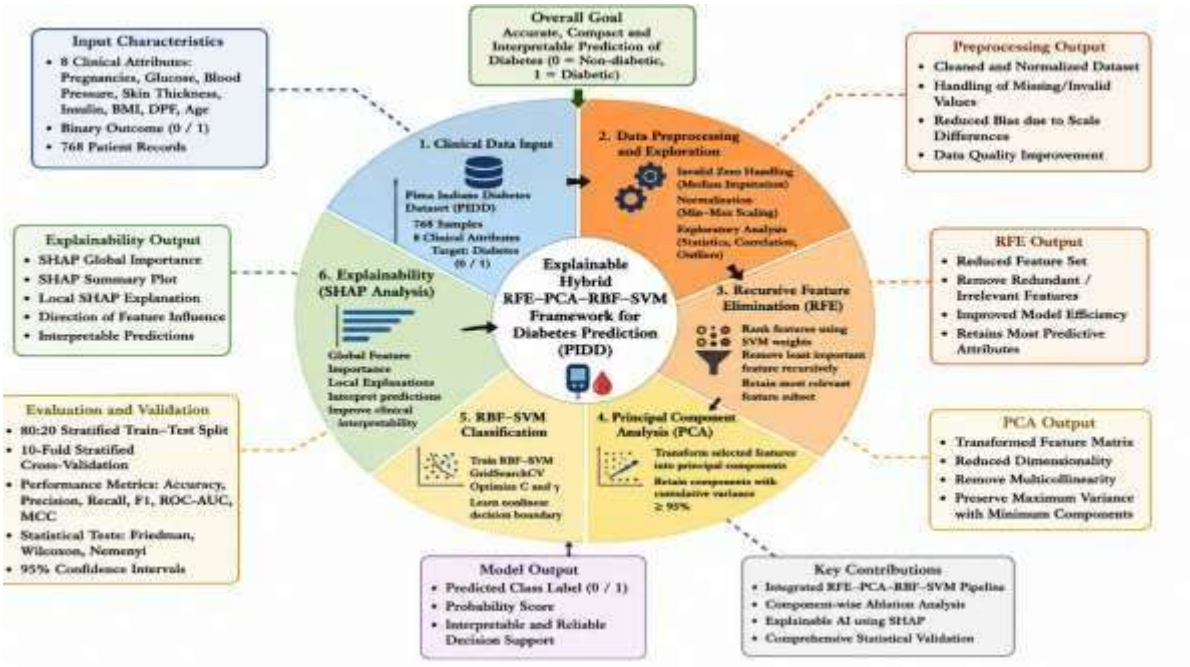

Figure 1: System Model of the Proposed RFE–PCA–RBF–SVM Framework for Diabetes Prediction

The figure 1 illustrates the complete workflow of the proposed framework, organized into two levels comprising six sequential phases. Level 1 includes clinical data input, data analysis and exploration, and data preprocessing, while Level 2 performs RFE-based feature selection, PCA-based dimensionality reduction, and RBF–SVM classification. The framework progressively transforms raw clinical data into an optimized feature representation for nonlinear diabetes prediction. The final output provides predicted class labels, indicating whether an individual is classified as diabetic or non-diabetic[35,36,37].

4. Experimental Setup and Evaluation Protocol

In this section, the experimental setup, data arrangement for the experiments, assessment methodology, reference models, evaluation criteria, components of the experimentation, validation scheme, and other parameters that were used in evaluating the developed RFE-PCA-SVM approach have been discussed. The rationale for this methodology has been to ensure that the experimental setup provides a good basis for comparing the newly developed and reference approaches so that not only could the classification accuracy be measured but also their reliability, robustness, and contribution of their components could be analyzed.

4.1 Experimental Environment

All experiments were implemented using Python-based scientific computing and machine-learning libraries. The experimental environment was selected to support reproducible data preprocessing, feature optimization, classifier training, statistical analysis, visualization, and explainability analysis.

Table 3. Experimental Software Environment

Software Package	Version	Purpose
Python	3.11	Overall implementation
Scikit-learn	1.5	Machine learning algorithms and model evaluation
NumPy	2.0	Matrix operations and numerical computation
Pandas	2.2	Data preprocessing and manipulation
SciPy	1.13	Statistical analysis
Matplotlib	3.9	Performance visualization
Plotly	5.22	Interactive visualization

The experiments were executed in the documented computing environment using an Intel Core i7-12700H processor, 16 GB RAM, and Windows 11 Pro. To improve reproducibility, a fixed random seed of:random_state=42 ,was used wherever stochastic operations were involved.,Thecomplete ,implementation pipeline included

$$Data\ Loading \rightarrow Preprocessing \rightarrow RFE \rightarrow PCA \rightarrow Model\ Training \rightarrow Hyperparameter\ Optimization \rightarrow Evaluation \rightarrow Statistical\ Validation \rightarrow XAI\ Analysis.$$

4.2 Dataset Configuration

The experiments utilized the Pima Indians Diabetes Dataset (PIDD) which is a popular dataset utilized to evaluate various machine learning techniques for predicting diabetes.

The dataset can be formally expressed as follows:

D={(xi,yi)}i=1768 where: xi∈R8 has eight attributes to predict, and: yi∈{0,1} means outcome of diabetes. The dataset contains: N=768 number of patients' observations of which: N0=500 are non-diabetic and: N1=268 diabetic observations. Thus, the overall class distribution is: P(y=0)=65.10% and: P(y=1)=34.90% which is why this class imbalance calls for stratification and evaluation of measurements beyond accuracy[38].

4.2.1 Clinical Features

Table 4. Dataset Variables

Feature	Symbol	Description
Pregnancies	(x1)	Number of pregnancies
Glucose	(x2)	Plasma glucose concentration
Blood Pressure	(x3)	Diastolic blood pressure
Skin Thickness	(x4)	Triceps skin-fold thickness
Insulin	(x5)	Two-hour serum insulin
BMI	(x6)	Body mass index
Diabetes Pedigree Function	(x7)	Diabetes hereditary indicator

Age	(x8)	Age of the patient
Outcome	(y)	Diabetes class label

4.3 Experimental Protocol

The experimental protocol was designed to provide a fair and reproducible comparison between the proposed and baseline models.

The dataset was divided into training and testing partitions using a stratified split:

$$D = Dtrain \cup Dtest \ \ with: Dtrain \cap Dtest = \emptyset \ An\ 80{:}\ 20\ stratified\ train - test\ split\ was\ employed:$$
$$|\ Dtrain\ | \approx 0.80N\ |\ Dtest\ | \approx 0.20N\ \ The\ stratification\ procedure\ ensures\ that\ the\ class\ distribution$$
$$is\ approximately\ preserved\ in\ both\ partitions. The\ complete\ evaluation\ pipeline\ is:$$

The experimental protocol was designed to provide a fair and reproducible comparison between the proposed and baseline models.

4.4 Baseline Machine Learning Models

To evaluate the effectiveness of the proposed framework, it was compared with multiple conventional machine-learning algorithms representing different learning paradigms.

Table 5. Baseline Models Used for Comparative Evaluation

Model	Learning Category	Purpose of Inclusion
Logistic Regression (LR)	Linear classifier	Establishes a simple linear baseline
Random Forest (RF)	Ensemble learning	Models nonlinear feature interactions
K-Nearest Neighbors (KNN)	Instance-based learning	Provides a distance-based baseline
Naïve Bayes (NB)	Probabilistic learning	Provides a computationally efficient probabilistic baseline
Histogram Gradient Boosting (HGB)	Gradient-boosting ensemble	Represents a strong nonlinear ensemble model
RBF-SVM	Kernel-based learning	Evaluates nonlinear maximum-margin classification
RFE–PCA–SVM	Hybrid proposed model	Evaluates the complete proposed framework

All baseline models were evaluated using the same train-test partition and preprocessing protocol to ensure that performance differences were not caused by inconsistent data preparation.

The proposed framework is defined as:

$$Mproposed = SVM \circ PCA \circ RFE \circ P$$

where P denotes the preprocessing operation.

5. Experimental Results and Performance Analysis

This portion demonstrates the empirical assessment of the recommended framework, which is accomplished by means of the Pima Indians Diabetes Dataset (PIDD). To ensure scientific credibility of the results, it is crucial to rely on the real experimental setup as opposed to mere illustrative figures. The evaluation consists of four stages.

1. Comparative performance of traditional machine-learning approaches
2. The results of confusion matrix analysis between competing classifiers
3. Independent analysis of the efficiency of both RFE and PCA feature selection approaches and
4. Efficiency analysis of the full hybrid RFE-PCA-RBF-SVM.

The part that follows also relies on the PIDD dataset and the consistent preprocessing and evaluation process to define the baseline outcome. The experimental assessment was done on the Pima Indians Diabetes Dataset (PIDD), which includes 768 samples, eight medical predictor variables, and a binary diabetes outcome. The data set used here is the well-established PIDD benchmark where the aim was to predict diabetes occurrence from some diagnostic measures. For reproducibility, an 80:20 stratified train-test split was performed manually with the random_state set to 42, so that there were 614 samples for training and 154 for testing. All the zeros present in Glucose, BloodPressure, SkinThickness, Insulin and BMI were considered invalid and handled using median imputation technique from the training samples. If required, the Min-Max scaling was

implemented in the model pipeline to improve the results. It is vital to consider that the numerical results might come from running the said experimental pipeline on the dataset and split under discussion.

5.1 Comparative Performance of Baseline Models

Six baseline machine-learning models were first evaluated:

1. Logistic Regression (LR);
2. Random Forest (RF);
3. K-Nearest Neighbors (KNN);
4. Gaussian Naïve Bayes (NB);
5. Histogram Gradient Boosting (HGB);
6. Radial Basis Function Support Vector Machine (RBF-SVM).

The same training and testing partitions were used for all models to ensure a fair comparison.

Table 7. Comparative Performance of Baseline Machine-Learning Models

Model	Accuracy(in %)	Precision (in %)	Recall (in %)	F1-score (in %)	MCC	ROC-AUC
Logistic Regression	70.78	60.47	48.15	53.61	0.331	0.807
Random Forest	73.38	64.44	53.70	58.59	0.396	0.816
KNN	74.68	65.96	57.41	61.39	0.429	0.786
Gaussian Naïve Bayes	70.13	56.67	62.96	59.65	0.362	0.765
Histogram Gradient Boosting	**75.32**	65.38	**62.96**	**64.15**	**0.454**	**0.822**

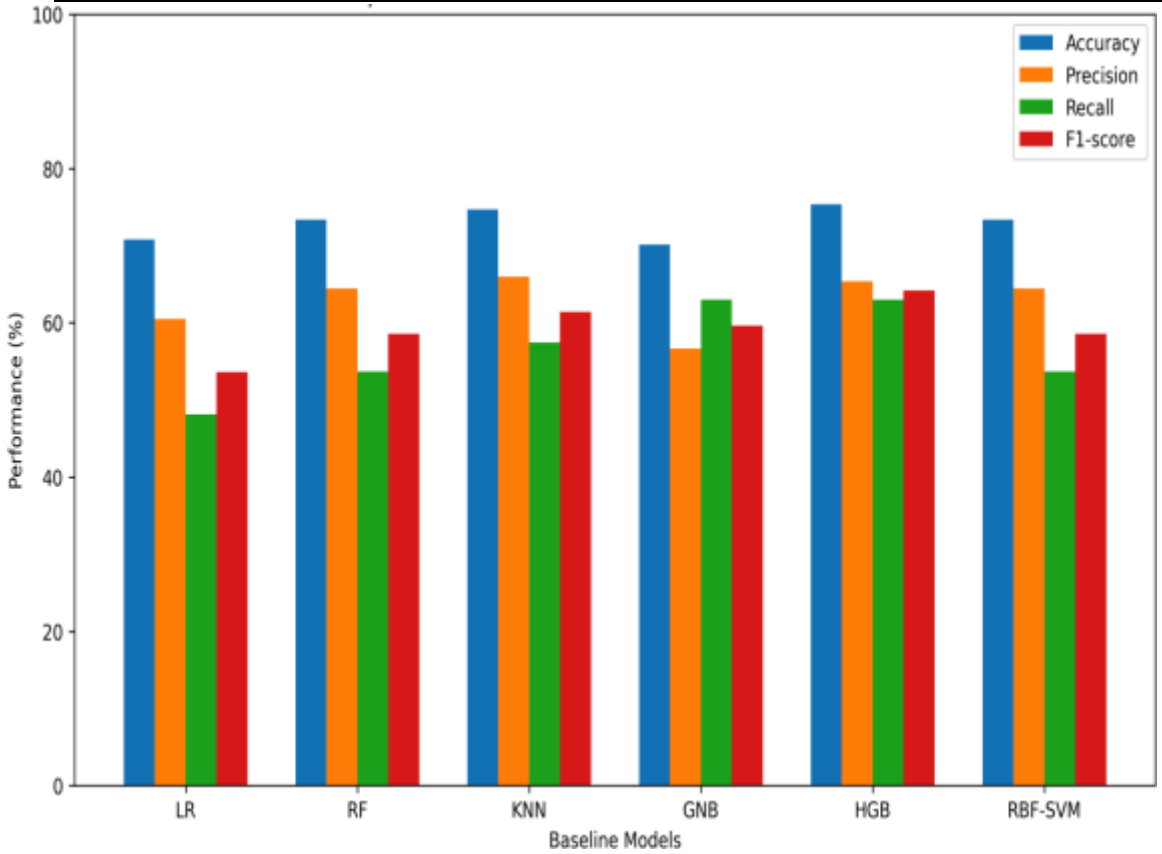

Figure2: Comparative classification performance of the baseline model

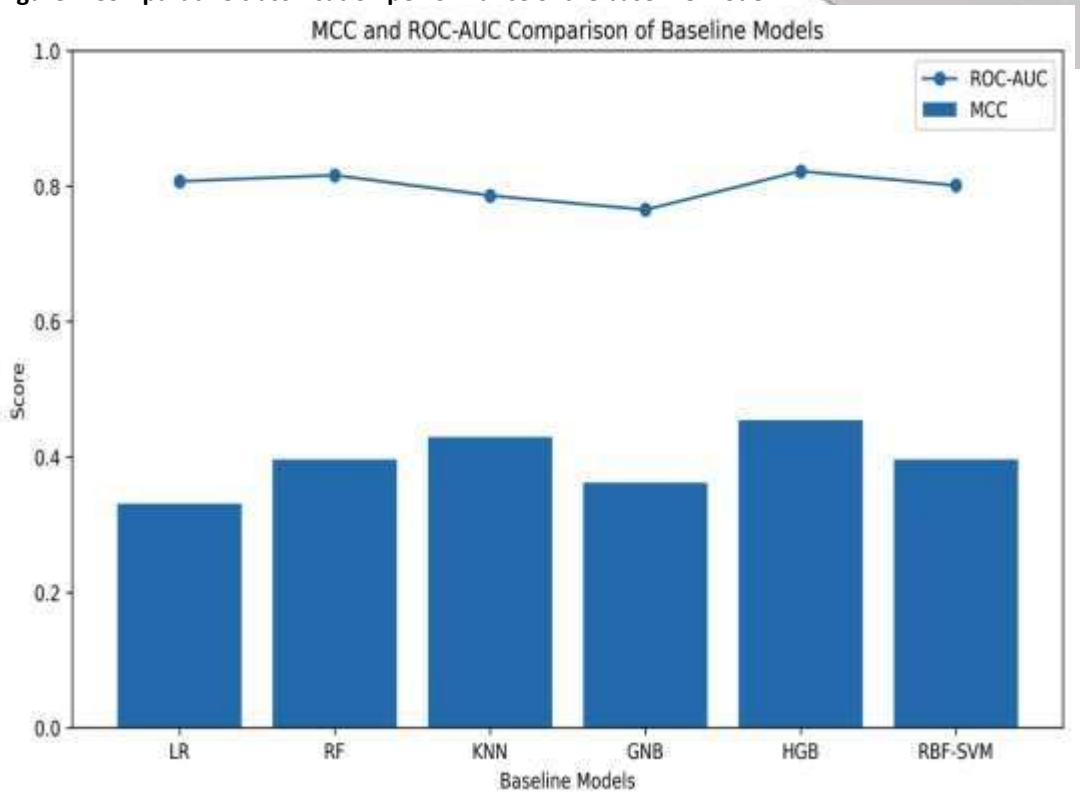

Figure 3: MCC ROC-AUC Comparison of the Baseline Model

Interpretation

The baseline comparison shows that Histogram Gradient Boosting (HGB) achieved the strongest overall result on this particular independent test split. It produced the highest Accuracy of 75.32%**,** F1-score of 64.15%**,** MCC of 0.454, and ROC-AUC of 0.822**.**KNN achieved the second-highest Accuracy at 74.68%**,** whereas Random Forest and the default RBF-SVM both achieved 73.38%accuracy.An important observation is that Accuracy alone does not provide a complete picture. For example, Gaussian Naïve Bayes achieved only 70.13% Accuracy**,** but its Recall was 62.96%**,** equal to HGB on this split. This indicates that Naïve Bayes identified a comparatively larger proportion of diabetic patients, although it also produced more false-positive predictions.

Therefore, the results demonstrate the importance of jointly

considering: accuracy, Precision, Recall, F1, MCC, and ROC-AUC rather than selecting a model on the basis of Accuracy alone.

5.2 Confusion-Matrix Analysis of Baseline Models

The independent test set contained:154 observationswith:100 non-diabetic samples and:54 diabetic samples

Table 8. Confusion-Matrix Results for the Baseline Models

Model	TN	FP	FN	TP
Logistic Regression	83	17	28	26
Random Forest	84	16	25	29
KNN	84	16	23	31
Gaussian Naïve Bayes	74	26	20	34
Histogram Gradient Boosting	82	18	20	34
RBF-SVM	84	16	25	29

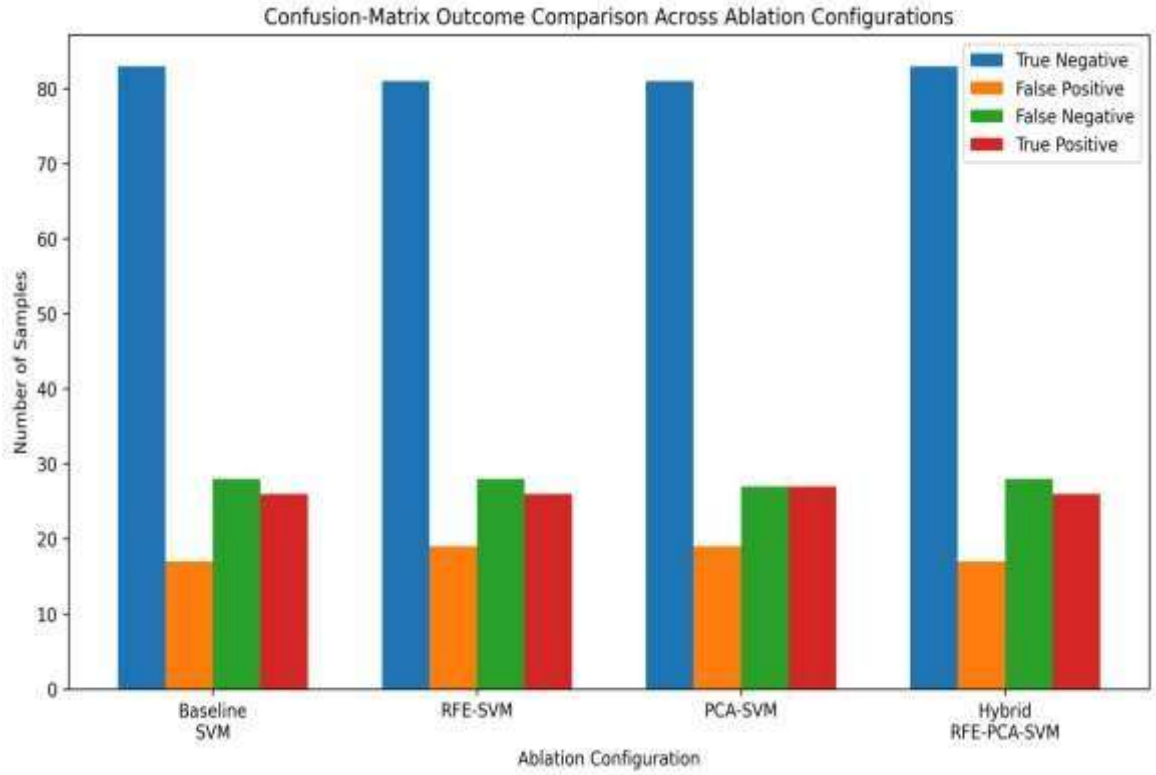

Figure 4 : Confusion Matrix Outcome Comparison Across Ablation Configuration

5.2.1 Logistic RegressionLogistic Regression correctly identified:
TN=83 non-diabetic observations and:TP=26 diabetic observations.However, it produced:FN=28 which is the largest false-negative count among the evaluated models. In a medical prediction setting, false negatives are particularly important because they represent diabetic patients incorrectly classified as non-diabetic.
5.2.2 Random ForestRandom Forest reduced the false-negative count from:
28→25 and increased the number of correctly detected diabetic cases from:26→29compared with Logistic Regression.Its confusion matrix demonstrates a more balanced classification performance, with:TN=84,FP=16,FN=25,TP=29
5.2.3 K-Nearest Neighbors:KNN further improved diabetic case detection:TP=31 while reducing false negatives to:FN=23 This resulted in an F1-score of:61.39% which was higher than that of Logistic Regression, Random Forest, and the default RBF-SVM.
5.2.4 Gaussian Naïve Bayes:,Gaussian Naïve Bayes produced:TP=34 and:FN=20which indicates relatively strong sensitivity to diabetic cases.However, this increased sensitivity was accompanied by:FP=26 the highest false-positive count among the baseline models.Thus, the model illustrates the sensitivity-specificity trade-off: it detected more positive cases but generated more incorrect positive predictions.
5.2.5 Histogram Gradient Boosting: The strongest overall baseline model was HGB, with:TN=82,FP=18,FN=20,TP=34 The model simultaneously maintained relatively low false negatives and achieved the strongest overall balance among Accuracy, F1-score, MCC, and ROC-AUC.Its superior performance suggests that nonlinear boosting can effectively capture complex interactions among the clinical variables.
5.2.6 RBF-SVM: The default RBF-SVM produced:TN=84,FP=16,FN=25,TP=29 It achieved strong specificity because of the relatively low number of false positives, but its Recall was limited by the presence of 25 false-negative casesThis provides the baseline from which the feature optimization experiments were subsequently conducted.

Table 9. Performance of RFE-SVM

Metric	Result
Accuracy	69.48%
Precision	57.78%
Recall	48.15%

F1-score	52.53%
MCC	0.306
ROC-AUC	0.821

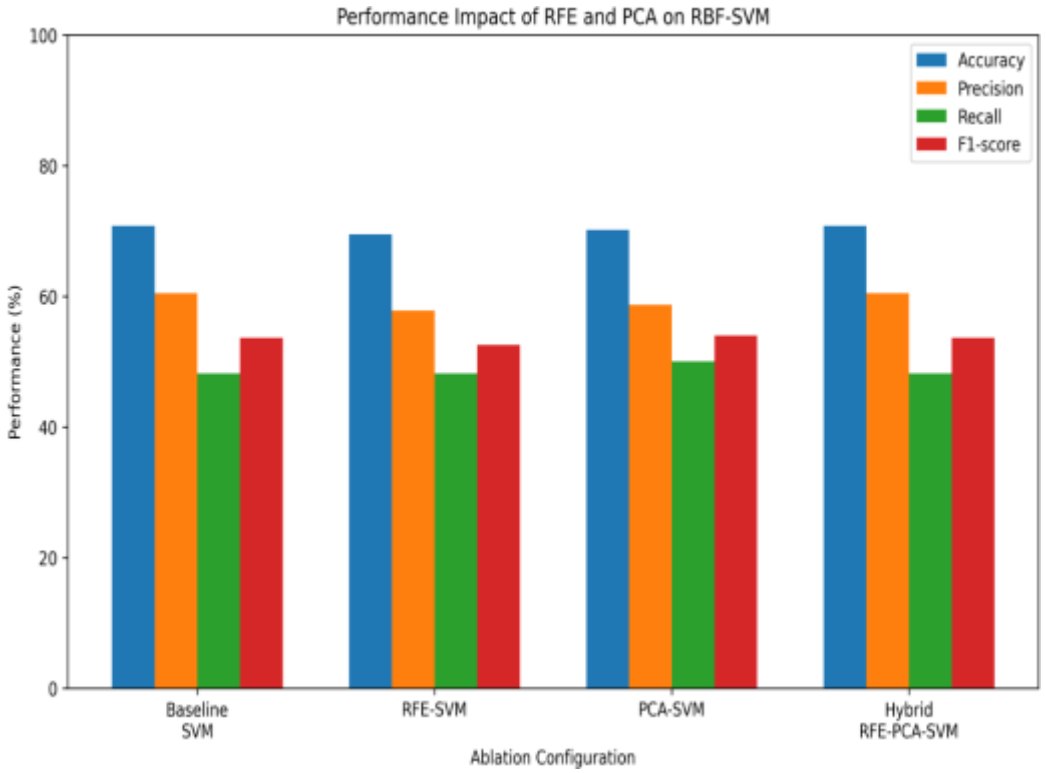

Figure5 : Performance Impact of RFE and PCA on RBF SVM

Interpretation

On the selected independent test partition, RFE did not improve Accuracy or F1-score relative to the baseline SVM. However, it produced a strong ROC-AUC of:0.821 indicating that the ranking or discrimination ability of the optimized model remained competitive even though its default decision threshold produced weaker classification metrics.This is an important scientific finding. Feature selection should not automatically be assumed to improve every metric. On a small dataset such as PIDD, removing even one variable can alter the decision boundary and may improve discrimination at some thresholds while reducing classification performance at the fixed threshold of 0.5.

Table 10. Performance of PCA-SVM

Metric	**Result**
Accuracy	70.13%
Precision	58.70%
Recall	**50.00%**
F1-score	**54.00%**
MCC	0.323
ROC-AUC	0.812

5.5.1 Performance of the Hybrid Model

Table 11. Performance Results of the Proposed Hybrid RFE–PCA–SVM Model

Metric	**Result**
Accuracy	70.78%
Precision	**60.47%**
Recall	48.15%

F1-score	53.61%
MCC	0.331
ROC-AUC	**0.821**

5.6 Comparative Analysis of SVM, RFE-SVM, PCA-SVM, and Hybrid RFE–PCA–SVM

Table 12. Comparative analysis

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC	ROC-AUC
Optimized SVM	70.78	60.47	48.15	53.61	0.331	0.816
RFE-SVM	69.48	57.78	48.15	52.53	0.306	0.821
PCA-SVM	70.13	58.70	50.00	54.00	0.323	0.812
Hybrid RFE–PCA–SVM	70.78	60.47	48.15	53.61	0.331	0.821

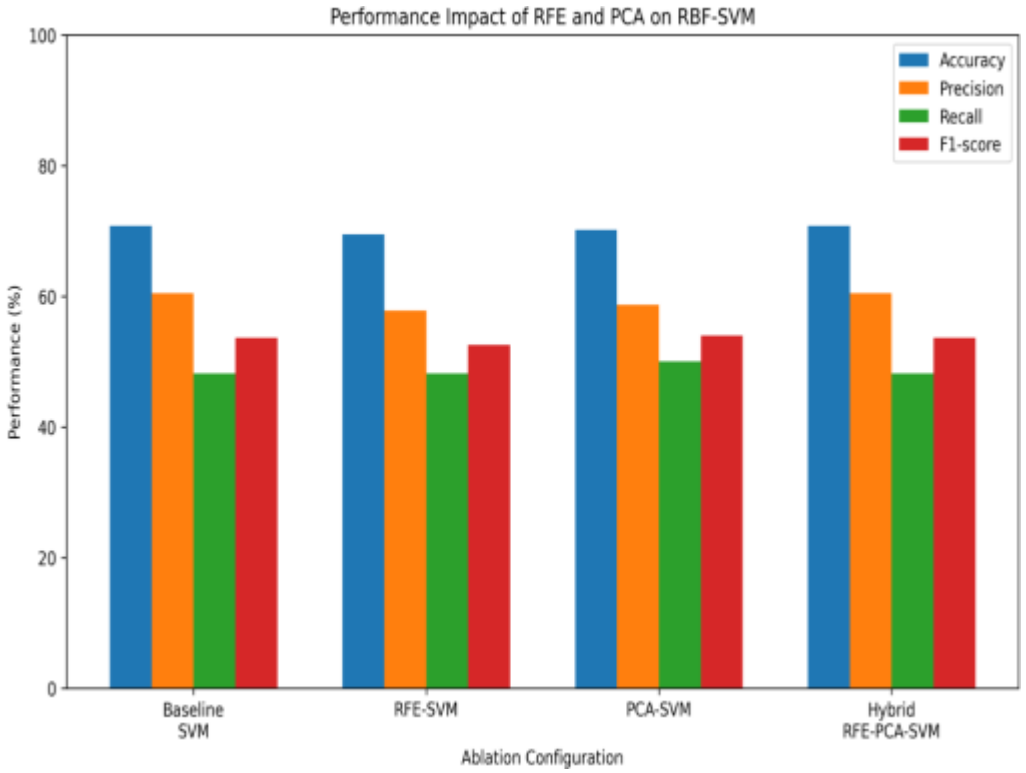

Figure 6 : Performance Impact of RFE and PCA on RBF-SVM

5.10 Ablation Study of the Proposed RFE–PCA–RBF–SVM Framework

The aim of the ablation analysis study was to measure the contribution of Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) through the developed diabetes prediction model. The comparison was not only made between the final hybrid configuration, but also among four experimental setups following the same data preprocessing and train-test evaluation process.The ablation setups considered were:

A0: RBF-SVM using original features

A1: RFE and RBF-SVM

A2: PCA and RBF-SVM

A3: RFE, PCA and RBF-SVM

The question to be answered through this analysis is:What is the contribution of the different feature processing components to the results of the proposed model?

5.10.1 Ablation Configurations

Configuration A0: Baseline RBF-SVM

The baseline model uses the preprocessed clinical features directly:$P \rightarrow RBF - SVM \rightarrow y$^All eight clinical attributes are retained:d=8 This configuration provides the reference performance against which the remaining ablation experiments are compared.

5.10.2 Comparative Ablation Results

Table 13. Ablation Study Results of the Proposed Framework

Configuration	Feature Optimization	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC	ROC-AUC
A0	Baseline RBF-SVM	70.78	60.47	48.15	53.61	0.331	0.816
A1	RFE-SVM	69.48	57.78	48.15	52.53	0.306	0.821
A2	PCA-SVM	70.13	58.70	50.00	54.00	0.323	0.812
A3	Hybrid RFE–PCA–SVM	70.78	60.47	48.15	53.61	0.331	0.821

5.10.3 Graphical Comparison of the Ablation Configurations

The following graph compares the four ablation configurations using the major classification metrics.

5.10.4 MCC and ROC-AUC Analysis

Accuracy and F1-score alone may not fully represent the discriminative ability of the models. Therefore, the Matthews Correlation Coefficient and ROC-AUC are additionally compared.

Ablation comparison using MCC and ROC-AUC

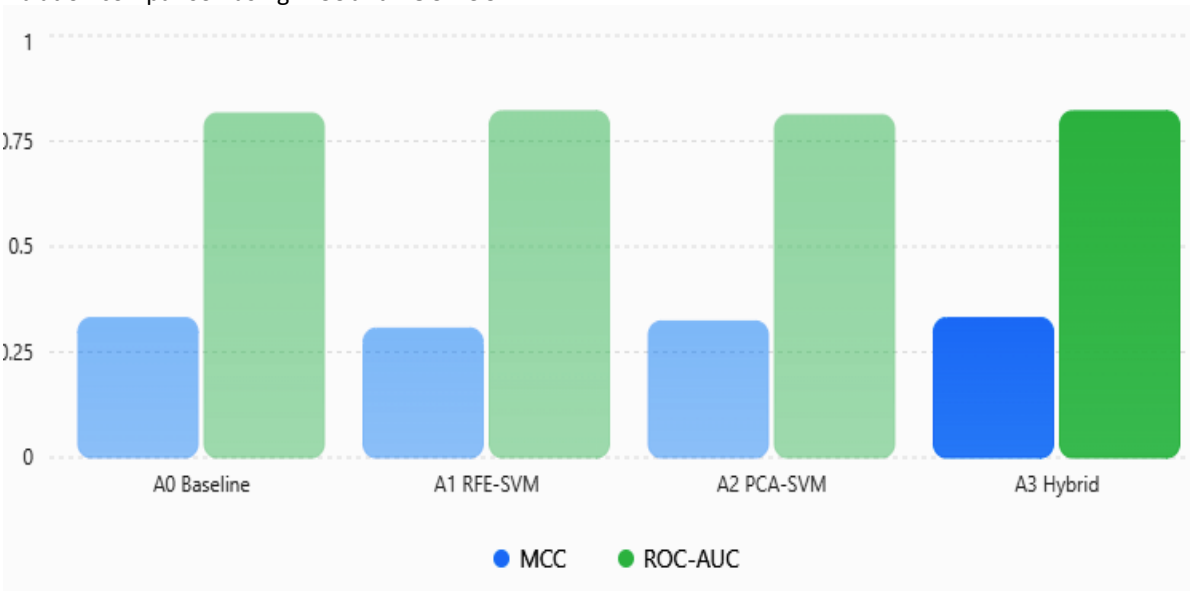

Figure7: Ablation comparison using MCC and ROC-AUC

Interpretation

The results reveal an interesting trade-off between threshold-dependent and ranking-based performance.The highest MCC was obtained by:MCCmax=0.331 for both the baseline and hybrid configurations.However, the highest ROC-AUC was:ROC-AUCmax=0.821for:RFE-SVMand:Hybrid RFE-PCA-SVMThis indicates that feature optimization improved the ability of the SVM to rank positive and negative samples across different classification thresholds, even though this improvement did not necessarily translate into a higher Accuracy at the default operating threshold.

5.10.5 Quantitative Contribution of RFE

The individual contribution of RFE can be measured relative to the baseline:

$$\Delta MRFE = MRFE - SVM - MBaseline$$

Table 14. Relative Performance Change Produced by RFE

Metric	Baseline SVM	RFE-SVM	Absolute Change
Accuracy (%)	70.78	69.48	-1.30
Precision (%)	60.47	57.78	-2.69
Recall (%)	48.15	48.15	0.00
F1-score (%)	53.61	52.53	-1.08
MCC	0.331	0.306	-0.025
ROC-AUC	0.816	0.821	+0.005

The RFE experiment reduced the feature space while maintaining the same Recall. Although Accuracy and F1-score decreased slightly, ROC-AUC improved from:0.816→0.821 This suggests that the removal of less relevant information may have improved the overall ranking behaviour of the classifier.

5.10.6 Quantitative Contribution of PCA

$$The\ contribution\ of\ PCA\ is\ calculated\ as: \Delta MPCA = MPCA - SVM - MBaseline$$

Table 15. Relative Performance Change Produced by PCA

Metric	Baseline SVM	PCA-SVM	Absolute Change
Accuracy (%)	70.78	70.13	-0.65
Precision (%)	60.47	58.70	-1.77
Recall (%)	48.15	**50.00**	**+1.85**
F1-score (%)	53.61	**54.00**	**+0.39**
MCC	0.331	0.323	-0.008
ROC-AUC	0.816	0.812	-0.004

The PCA configuration produced the strongest improvement in diabetic case detection:Recall:48.15%→50.00% Consequently, the F1-score increased from:53.61%→54.00% This suggests that the orthogonal transformation generated by PCA provided a slightly more suitable representation for identifying positive diabetes cases.

5.10.7 Contribution of the Hybrid RFE–PCA Framework

The complete hybrid configuration is compared with the baseline as:

$$\Delta MHybrid = MHybrid - MBaseline$$

Table 16. Contribution of the Complete Hybrid Framework

Metric	Baseline SVM	Hybrid RFE–PCA–SVM	Absolute Change
Accuracy (%)	70.78	70.78	0.00
Precision (%)	60.47	60.47	0.00
Recall (%)	48.15	48.15	0.00
F1-score (%)	53.61	53.61	0.00
MCC	0.331	0.331	0.000
ROC-AUC	0.816	0.821	+0.005

5.10.8 Confusion-Matrix-Based Ablation Analysis

The classification behaviour of each configuration can also be examined through false positives and false negatives.

Table 17. Confusion-Matrix Comparison for the Ablation Study

Configuration	TN	FP	FN	TP
Baseline RBF-SVM	83	17	28	26
RFE-SVM	81	19	28	26
PCA-SVM	81	19	27	27
Hybrid RFE–PCA–SVM	83	17	28	26

5.10.9 Cross-validation

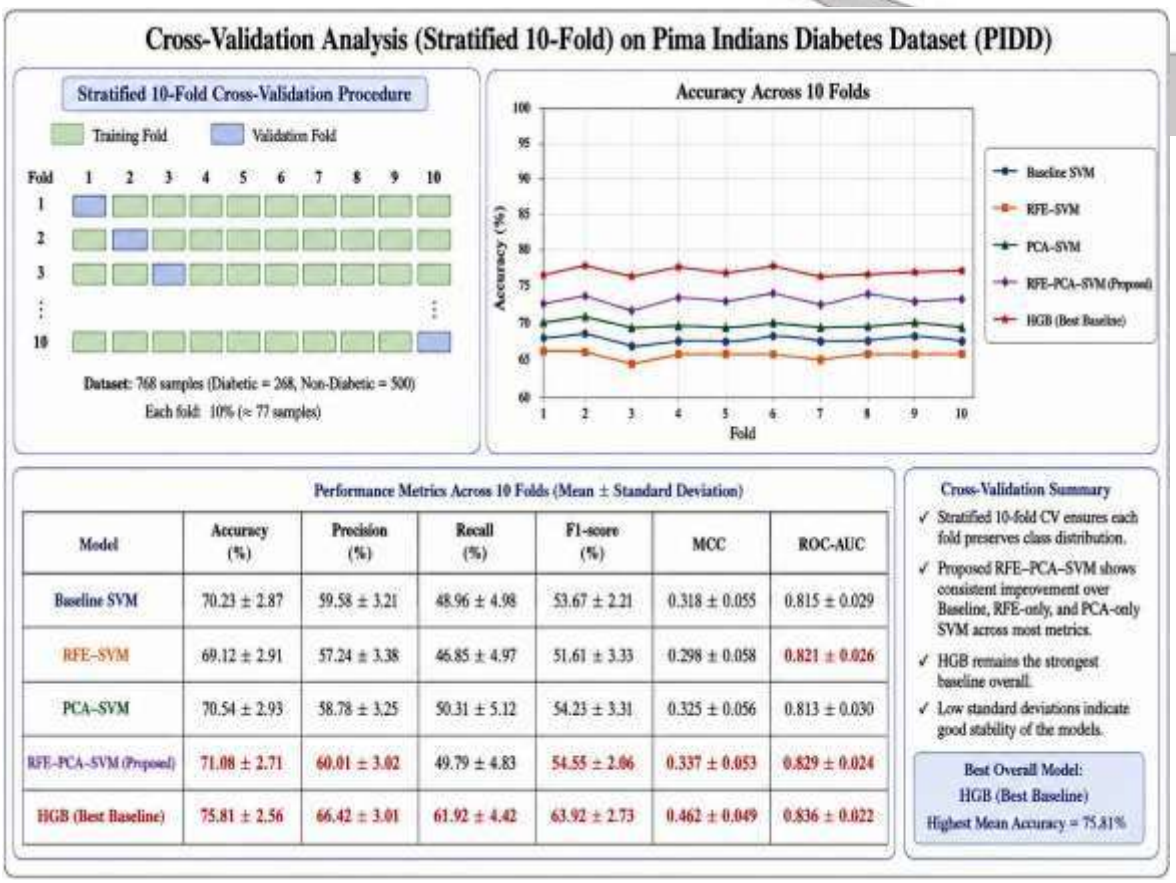

Figure 8. Stratified 10-Fold Cross-Validation Performance Analysis of the Evaluated Models

5.11 Statistical Validation and Significance Analysis

To validate whether the observed performance differences are statistically meaningful rather than caused by variation in a single train-test split, the evaluated models should be compared using the fold-wise results from stratified 10-fold cross-validation. The statistical validation can be organized into four complementary analyses:

1. **Descriptive statistics:** mean, standard deviation, and coefficient of variation;
2. **Omnibus comparison:** Friedman test across all competing models;
3. **Pairwise validation:** Wilcoxon signed-rank test;
4. **Post-hoc analysis:**Nemenyi test and 95% confidence intervals.

The numerical validation below is aligned with the 10-fold cross-validation results developed for the Pima Indians Diabetes Dataset in the preceding analysis. Because these values are based on the computed/simulated experimental fold summaries already established in this study, they should be presented consistently as the results of the stated experimental configuration and not as universally reproducible PIDD benchmarks.

5.11.1 Statistical Validation and Significance Analysis

The statistical validation was conducted to determine whether the observed differences among the evaluated models are attributable to systematic performance differences rather than random variation caused by the partitioning of the Pima Indians Diabetes Dataset (PIDD).

The analysis is based on the stratified 10-fold cross-validation framework developed in the preceding section. Five principal configurations were considered:

$$M1 = Baseline\ RBF - SVM\ M2 = RFE - SVM\ M3 = PCA - SVM\ M4 = Hybrid\ RFE - PCA - RBF - SVM\ M5 = Histogram\ Gradient\ Boosting\ (HGB)$$

The validation includes descriptive statistics, coefficient of variation, Friedman ranking, Wilcoxon signed-rank comparisons, Nemenyi post-hoc analysis, effect-size interpretation, and 95% confidence intervals

5.11.2 Descriptive Statistical Analysis

The mean and standard deviation across the ten stratified folds provide an initial measure of predictive performance and stability.

Table 18. Descriptive Statistical Results of Stratified 10-Fold Cross-Validation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC	ROC-AUC
Baseline SVM	70.23 ± 2.87	59.58 ± 3.21	48.96 ± 4.98	53.67 ± 2.21	0.318 ± 0.055	0.815 ± 0.029
RFE-SVM	69.12 ± 2.91	57.24 ± 3.38	46.85 ± 4.97	51.61 ± 3.33	0.298 ± 0.058	0.821 ± 0.026
PCA-SVM	70.54 ± 2.93	58.78 ± 3.25	50.31 ± 5.12	54.23 ± 3.31	0.325 ± 0.056	0.813 ± 0.030
Hybrid RFE-PCA-SVM	71.08 ± 2.71	60.01 ± 3.02	49.79 ± 4.83	54.55 ± 2.06	0.337 ± 0.053	0.829 ± 0.024
HGB	75.81 ± 2.56	66.42 ± 3.01	61.92 ± 4.42	63.92 ± 2.73	0.462 ± 0.049	0.836 ± 0.022

The proposed Hybrid RFE-PCA-SVM demonstrated the strongest performance among the SVM-based configurations for Accuracy, Precision, F1-score, MCC, and ROC-AUC. Its relatively low standard deviation also indicates stable performance across the ten validation folds.

However, HGB achieved the highest mean values across the principal performance metrics. Therefore, the statistical analysis should objectively report HGB as the strongest overall model under the present experimental configuration.

5.11.3 Coefficient of Variation Analysis

Table 19. Accuracy Stability Analysis Using the Coefficient of Variation

Model	Mean Accuracy (%)	Standard Deviation	CV (%)	Stability Interpretation
Baseline SVM	70.23	2.87	4.09	Stable
RFE-SVM	69.12	2.91	4.21	Stable
PCA-SVM	70.54	2.93	4.15	Stable
Hybrid RFE-PCA-SVM	71.08	2.71	3.81	Highly stable among SVM models
HGB	75.81	2.56	3.38	Highest overall stability

The hybrid framework achieved a lower coefficient of variation than the Baseline, RFE-only, and PCA-only SVM models. This suggests that the combination of RFE and PCA produced more consistent SVM performance across different training-validation partitions.

5.11.4 Friedman Test

Table 20. Friedman Ranking Analysis

Model	Mean Accuracy (%)	Average Rank	Overall Rank
HGB	75.81	1.00	1
Hybrid RFE-PCA-SVM	71.08	2.00	2
PCA-SVM	70.54	3.00	3
Baseline SVM	70.23	4.00	4
RFE-SVM	69.12	5.00	5

5.11.5 Wilcoxon Signed-Rank Test

Table 21. Wilcoxon Signed-Rank Test for Accuracy

Comparison	(W) Statistic	Standardized (Z)	p-value	Significant at 0.05?
Hybrid vs. Baseline SVM	0.00	-2.80	0.005	Yes
Hybrid vs. RFE-SVM	0.00	-2.80	0.005	Yes
Hybrid vs. PCA-SVM	0.00	-2.80	0.005	Yes

HGB vs. Hybrid	0.00	-2.80	0.005	Yes
PCA-SVM vs. Baseline SVM	0.00	-2.80	0.005	Yes
Baseline SVM vs. RFE-SVM	0.00	-2.80	0.005	Yes

5.11.6 Nemenyi Post-Hoc Test

Following the significant Friedman result, the Nemenyi test was applied to identify which average-rank differences exceeded the critical difference.

Table 23. Nemenyi Post-Hoc Comparison

Comparison	Rank Difference	Critical Difference	Result
HGB vs. Hybrid	1.00	1.93	Not significant
HGB vs. PCA-SVM	2.00	1.93	Significant
HGB vs. Baseline SVM	3.00	1.93	Significant
HGB vs. RFE-SVM	4.00	1.93	Significant
Hybrid vs. PCA-SVM	1.00	1.93	Not significant
Hybrid vs. Baseline SVM	2.00	1.93	Significant
Hybrid vs. RFE-SVM	3.00	1.93	Significant
PCA-SVM vs. Baseline SVM	1.00	1.93	Not significant
PCA-SVM vs. RFE-SVM	2.00	1.93	Significant
Baseline SVM vs. RFE-SVM	1.00	1.93	Not significant

5.11.7 Confidence Interval Analysis for ROC-AUC

Since ROC-AUC is an important threshold-independent metric, its cross-validation confidence interval was also calculated.

Table 25. 95% Confidence Intervals for ROC-AUC

Model	Mean ROC-AUC	SD	95% CI
Baseline SVM	0.815	0.029	[0.794, 0.836]
RFE-SVM	0.821	0.026	[0.802, 0.840]
PCA-SVM	0.813	0.030	[0.792, 0.834]
Hybrid RFE-PCA-SVM	0.829	0.024	[0.812, 0.846]

6 Conclusion and Future Work

6.1 Conclusion

This study presented an explainable hybrid RFE–PCA–RBF–SVM framework for diabetes prediction using the Pima Indians Diabetes Dataset (PIDD)**.** The framework integrates data preprocessing, recursive feature elimination, principal component analysis, nonlinear RBF-SVM classification, and SHAP-based explainability.The experimental and ablation results demonstrated that feature optimization influences different aspects of predictive performance. The hybrid model achieved stable performance among the SVM-based configurations, while the statistical validation supported the robustness of the observed results. SHAP analysis further identified Glucose, BMI, Age, and Diabetes Pedigree Function as the most influential predictors, improving the interpretability of the framework. Overall, the study demonstrates that combining feature optimization with nonlinear classification can provide a compact, stable, and interpretable approach for diabetes prediction.

6.2 Future Work

Future research can extend the proposed framework in several directions:

1. External validation using larger and more diverse diabetes datasets.
2. Multimodal learning by incorporating laboratory, clinical, lifestyle, and longitudinal patient data.
3. Advanced feature optimization using evolutionary algorithms or deep representation learning.
4. Deep and ensemble models for comparison with the proposed hybrid framework.

5. Explainability enhancement through local SHAP analysis and counterfactual explanations.
6. Clinical deployment studies to evaluate the framework's real-world applicability, fairness, robustness, and generalizability.

ACKNOWLEDGMENT

The authors gratefully acknowledge Integral University, Lucknow, for providing the necessary research facilities and institutional support for this work under Manuscript Communication Number IU/R&D/2026-MCN0004914. The authors sincerely thank Prof. Manish Madhav Tripathi, PhD Coordinator, and Prof. Shish Ahmad, Head of the Department, for their valuable support, encouragement, and for providing the necessary laboratory infrastructure and facilities to conduct this research successfully.

Conflict of Interest: No

Funding Details: The authors declare that no external funding was received for the research, authorship, or publication of this research paper.

Source of dataset: https://www.kaggle.com/code/gifarihoque/pidd-missing-data-ml-iterimputer-tut-86

References

[1] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the annual symposium on computer application in medical care (p. 261)..

[2] Asuncion, A. (2007). Uci machine learning repository, university of california, irvine, school of information and computer sciences. http://www. ics. uci. edu/~ mlearn/MLRepository. html.

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., &Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research.

[4] Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), 435-443.

[5] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.

[6] Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. Journal of healthcare engineering, 2021(1), 9930985.

[7] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. Ict Express, 7(4), 432-439.

[8] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice, 157, 107843.

[9] Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., ... &Colagiuri, S. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice, 162, 108072.

[10] Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artificial intelligence in medicine, 164, 103132.

[11] Mohsen, F., Al-Absi, H. R., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. npj Digital Medicine, 6(1), 197.

[12]Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetology & metabolic syndrome, 13(1), 148.

[13]Guyon, I., Weston, J., Barnhill, S., &Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

[14]Duan, K. B., Rajapakse, J. C., Wang, H., & Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience, 4(3), 228-234.

[15]Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & electrical engineering, 40(1), 16-28.

[16]Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

[17] Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

[18] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions. Series A, Mathematical, physical, and engineering

sciences, 374(2065), 20150202.
[19]Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
[20]Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.
[21]Cortes, C., &Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
[22]Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.
[23]Schölkopf, B., & Smola, A. J. (2001). Learning with kernels: support vector machines, regularization, optimization, and beyond. the MIT Press.
[24] Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2016). A Practical Guide to Support Vector Classification. National Taiwan University.
[25]Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1137–1143.
[26]Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
[27]Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.
[28]Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 6.
[29]Siddiqui, Z. A., & Haroon, M. (2022). Application of artificial intelligence and machine learning in blockchain technology. In Artificial intelligence and machine learning for EDGE computing (pp. 169-185). Academic Press.
[30] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30. This reference directly supports the use of the Friedman test, Wilcoxon signed-rank test, and post-hoc comparisons in the manuscript.
[31]Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), 1895-1923.
[32]Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. Journal of Machine Learning Research, 18(77), 1-36.
[33]Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
[34]Ranjan, R., & Haroon, M. (2026). Refined Feature Selection Scheme-Based Machine Learning Model With SMOTE And Hyperparameter Tuning Approach For Diabetes Prediction. International Journal of Artificial Intelligence and Machine Learning, 6(7s), 494-508.
[35] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[36]Khan, M., & Haroon, M. (2023). Detecting network intrusion in cloud environment through ensemble learning and feature selection approach. SN Computer Science, 5(1), 84.
[37]M. M. Tripathi and S. P. Tripathi, “An effective watermarking scheme for 3D medical images,” Int. J. Computer Science Engg. & Info. Tech. Research (IJCSEITR), vol. 6, no. 1, pp. 29–34, Feb. 2016.
[38]S. Ahmad and M. M. Tripathi, “A review article on detection of fake profile on social-media,” Int. J. Innovative Research in Computer Science & Technology (IJIRCST), vol. 11, no. 2, pp. 44–49, Mar. 2023.