ISSN:1533-9211
SEYBOLD Journal REPORT
www.seybold-report.com

Enhancing Diabetes Prediction Accuracy with AdvanSVM: A Machine Learning Approach Using the PIMA Dataset

Asha V [1], Manjunath Ramanna Lamani[2],Padmaja K[3] , Tejashwini A Gondhale[4]

[1]Research Scholar, RNS Institute of Technology, Bengaluru
[2]Research Scholar, CHRIST (Deemed to be University),Bengaluru
[3]Assistant Professor, GSSS, Mysore
[4]Assistant Professor, Computer Science Department, Dayananda Sagar College of Engineering, Bengaluru, Karnataka.

Abstract

Current research introduces the AdvanSVM, an improved Support Vector Machine designed to refine diabetes predictions using the PIMA Indian Diabetes dataset. The study explores various imputation methods to assess their impact on predictive precision, underscoring the significance of feature selection for fine-tuning machine learning models targeting diabetes forecasts. It also points out the importance of data balance and sophisticated model formulation, along with the need for broader datasets to propel forward strides in this field. Challenges pertaining to the specificity of the dataset and the extension of the results beyond its scope are acknowledged. TheAdvanSVM classifier, featuring custom kernel functions and tailored parameter adjustments, addresses these issues while also managing the inherent imbalance found in medical data. The research targets the unique challenges of the PIMA dataset, such as missing information and anomalies, striving for improved, clinically applicable predictions of diabetes. A thorough evaluation of the model with relevant metrics confirms its potential for providing precise diabetes predictions.

ISSN:1533-9211
SEYBOLD Journal REPORT
www.seybold-report.com

ISSN: 1533 - 9211

CORRESPONDING AUTHOR:

Asha V
asha.shekar91@gmail.com

KEYWORDS:

Diabetes Prediction, Machine Learning, Support Vector Machine, AdvanSVM, PIMAIndian Diabetes Dataset,

Received: 19 April 2024
Accepted: 10 May 2024
Published: 19 May 2024

TO CITE THIS ARTICLE:

Asha, V., Lamani, M. R., Padmaja, K., & Gondhale, T. A. (2024). Enhancing diabetes prediction accuracy with AdvanSVM: A machine learning approach using the PIMA dataset. *Seybold Report Journal*, *19*(05), 51-72. DOI: 10.5110/77. 1412

Introduction

In the evolving domain of healthcare informatics, the methodical examination of data imputation strategies is vital for enhancing classification tasks. Our study delves into an in-depth analysis of various imputation methods within the PIMA Indian Diabetes dataset, aiming to elevate diabetes classification accuracy. By evaluating Multivariate Imputation by Chained Equations (MICE), k-Nearest Neighbours (KNN), Mean, and Median imputation, we assess their impact on the fidelity of machine learning models in diabetes prediction[1].

In recent years, data mining and machine learning methods in the medical field have received much attention and have optimized many complex issues in the medical field. One of the problems facing researchers is the appropriate dataset, and the suitable dataset on which different methods of data mining and machine learning can be applied is rarely found. One of the most reliable and appropriate datasets in the field of diabetes diagnosis is the Indian Survey Database. In this article, we have tried to review the methods that have been implemented in recent years using machine learning classification algorithms on this data set and compare these methods in terms of evaluation criteria and feature selection methods. After comparing these methods, it was found that models that used feature selection methods were more accurate than other approaches[2].

Diabetes mellitus, a chronic metabolic disorder, continues to be a major public health issue around the world. It is estimated that one in every two diabetics is undiagnosed. Early diagnosis and management of diabetes can also prevent or delay the onset of complications. With the help of a variety of machine learning and deep learning models, stacking algorithms, and other techniques, our study’s goal is to detect diseases early. In this study, we propose two stackingbased models for diabetes disease classification using a combination of the PIMA Indian diabetes dataset, simulated data, and additional data collected from a local healthcare facility. We use both the classical and deep neural network stacking ensemble methods to combine the predictions of multiple classification models and improve classification accuracy and robustness. In the evaluation protocol, we used both the train-test and cross-validation (CV) techniques to validate our proposed model. The highest accuracy is obtained by stacking ensemble with three NN architectures, resulting in an accuracy of 95.50 %, precision of 94 %, recall of 97 %, and f1- score of 96 % using 5-fold CV on simulation study. The stacked accuracy obtained from ML algorithms for the Pima Indian Diabetes dataset is 75.03 % using the train-test split protocol, while the accuracy obtained from the CV protocol is 77.10 % on the stacked model. The range of performance scores that outperformed the CV protocol 2.23 %–12 %. Our proposed method achieves a high accuracy range from 92 % to 95 %, precision, recall, and F1-score ranges from 88 % to 96 % using classical and deep neural network (NN)-based stacking method on the primary dataset. The proposed dataset and ensemble method could be useful in the early detection and treatment of diabetes, as well as in the advancement of machine learning and data analysis techniques in the healthcare industry[3].

The Diabetes Mellitus (DM) is known as of the persistent disease which is due to excessive blood sugar levels. When it is left untreated it leads to severe health complications like cardiac disorders, kidney damage and stroke. The existing methods based on machine learning and deep learning approaches faces problem in predicting the diabetes of the patients in a precise manner. Moreover, the classification accuracy was diminished when evaluated for large datasets, so this research introduced an effective classification approach using the combination of Latent Dirichlet Allocation (LDA) and Artificial Neural Network (ANN). The probability distribution function of LDA is combined using the back propagation of ANN where the weights are initialized to perform an effective diabetes classification. The data is obtained from PIMA Dataset and North California State University (NCSU) dataset then the pre-processing is performed using min-max normalization approach. After this, Bivariate filter based feature selection is performed to choose appropriate features that were selected using the bivariate filter method is fed as input to Pearson correlation which selects the effective features based on threshold value. Finally, classification is performed using the proposed ANN-LDA. The results show that suggested method performs better than the existing approaches and achieves classification accuracy of 93%[4].

Diabetes is the most common disease and is a major cause for blindness, kidney failure, heart attacks, stroke and lower limb amputation. Thus, early prediction of diabetes is very crucial to initiating proper treatment to avoid further serious complications of the disease. The performance of recent diabetes detection schemes based on clinical data is highly influenced by low feature distinctiveness and unwanted features such as dermatologic manifestations. Different machine learning classifiers need tedious hyper-parameter tuning, which fails to assure a better diabetes detection rate. This article presents an analytical model to detect diabetes based on an optimized Support Vector Machine (SVM), K Nearest Neighbor (KNN) and Random Forest (RF) using decision level fusion to improve the diabetes detection rate. The hyper-parameters SVM, KNN, and RF are optimized using a multi-objective function-based Particle Swarm Optimization (PSO) algorithm, which considers various clinical entities for the diabetes detection, such as age, body mass index (BMI), blood pressure (BP), glucose, insulin, number of pregnancies, skin thickness, and diabetes pedigree function. The extensive experiments on the Indian Pima diabetes dataset confirmed that the diabetes detection using hybrid classifiers can provide a better prediction rate (94.27%) compared with single classifiers and the previous state of arts[5].

Sugar in the blood can harm individuals and their vital organs, potentially leading to blindness, renal illness, as well as kidney and heart diseases. Globally, diabetic patients face an average annual mortality rate of 38%. This study employs Chi-square, mutual information, and sequential feature selection (SFS) to choose features for training multiple classifiers. These classifiers include an artificial neural network (ANN), a random forest (RF), a gradient boosting (GB) algorithm, Tab-Net, and a support vector machine (SVM). The goal is to predict the onset of diabetes at an earlier age. The classifier, developed based on the selected features, aims to enable early diagnosis of diabetes. The PIMA and early-risk diabetes datasets serve as test subjects for the developed system. The feature selection technique is then applied to focus on the most important and relevant features for model training. The experiment findings conclude that the ANN exhibited a spectacular performance in terms of accuracy on the PIMA dataset, achieving a remarkable accuracy rate of 99.35%. The second experiment, conducted on the early diabetes risk dataset using selected features, revealed that RF achieved an accuracy of 99.36%. Based on our experimental results, it can be concluded that our suggested method significantly outperformed baseline machine learning algorithms already employed for diabetes prediction on both datasets[6].

Diabetes mellitus, usually called diabetes, is a serious public health issue that is spreading like an epidemic around the world. It is a condition that results in elevated glucose levels in the blood. India is often referred to as the 'Diabetes Capital of the World', due to the country's 17% share of the global diabetes population. It is estimated that 77 million Indians over the age of 18 have diabetes (i.e., everyone in eleven) and there are also an estimated 25 million pre-diabetics. One of the solutions to control diabetes growth is to detect it at an early stage which can lead to improved treatment. So, in this project, we are using a few machine learning algorithms like SVM, Decision Tree Classifier, Random Forest, KNN, Linear regression, Logistic regression, Naive Bayes to effectively predict the diabetes. Pima Indians Diabetes Database has been used in this project. According to the experimental findings, Random Forest produced an accuracy of 91.10% which is higher among the different algorithms used[7].

Diabetes, resulting from inadequate insulin production or utilization, causes extensive harm to the body. Existing diagnostic methods are often invasive and come with drawbacks, such as cost constraints. Although there are machine learning models like Classwise k Nearest Neighbor (CkNN) and General Regression Neural Network (GRNN), they struggle with imbalanced data and result in under-performance. Leveraging advancements in sensor technology and machine learning, we propose a non-invasive diabetes diagnosis using a Back Propagation Neural Network (BPNN) with batch normalization, incorporating data re-sampling and normalization for class balancing. Our method addresses existing challenges such as limited performance associated with traditional machine learning. Experimental results on three datasets show significant improvements in overall accuracy, sensitivity, and specificity compared to traditional methods. Notably, we achieve accuracies of 89.81% in Pima diabetes dataset, 75.49% in CDC BRFSS2015 dataset, and 95.28% in Mesra Diabetes dataset. This underscores the potential of deep learning models for robust diabetes diagnosis[8].

With the increasing prevalence of diabetes in Saudi Arabia, there is a critical need for early detection and prediction of the disease to prevent long-term health complications. This study addresses this need by using machine learning (ML) techniques applied to the Pima Indians dataset and private diabetes datasets through the implementation of a computerized system for predicting diabetes. In contrast to prior research, this study employs a semisupervised model combined with strong gradient boosting, effectively predicting diabetes-related features of the dataset. Additionally, the researchers employ the SMOTE technique to deal with the problem of imbalanced classes. Ten ML classification techniques, including logistic regression, random forest, KNN, decision tree, bagging, AdaBoost, XGBoost, voting, SVM, and Naive Bayes, are evaluated to determine the algorithm that produces the most accurate diabetes prediction. The proposed approach has achieved impressive performance. For the private dataset, the XGBoost algorithm with SMOTE achieved an accuracy of 97.4%, an F1 coefficient of 0.95, and an AUC of 0.87. For the combined datasets, it achieved an accuracy of 83.1%, an F1 coefficient of 0.76, and an AUC of 0.85. To understand how the model predicts the final results, an explainable AI technique using SHAP methods is implemented. Furthermore, the study demonstrates the adaptability of the proposed system by applying a domain adaptation method. To further enhance accessibility, a mobile app has been developed for instant diabetes prediction based on user-entered features. This study contributes novel insights and techniques to the field of ML-based diabetic prediction, potentially aiding in the early detection and management of diabetes in Saudi Arabia[9].

Diabetes prediction is an ongoing study topic in which medical specialists are attempting to forecast the condition with greater precision. Diabetes typically stays lethargic, and on the off chance that patients are determined to have another illness, like harm to the kidney vessels, issues with the retina of the eye, or a heart issue, it can cause metabolic problems and various complexities in the body. Various worldwide learning procedures, including casting a ballot, supporting, and sacking, have been applied in this review. The Engineered Minority Oversampling Procedure (Destroyed), along with the K-overlay cross-approval approach, was utilized to achieve class evening out and approve the discoveries. Pima Indian Diabetes (PID) dataset is accumulated from the UCI Machine Learning (UCI ML) store for this review, and this dataset was picked. A highlighted engineering technique was used to calculate the influence of lifestyle factors. A two-phase classification model has been developed to predict insulin resistance using the Sequential Minimal Optimisation (SMO) and SMOTE approaches together. The SMOTE technique is used to preprocess data in the model’s first phase, while SMO classes are used in the second phase. All other categorization techniques were outperformed by bagging decision trees in terms of Misclassification Error rate, Accuracy, Specificity, Precision, Recall, F1 measures, and ROC curve. The model was created using a combined SMOTE and SMO strategy, which achieved 99.07% correction with 0.1 ms of runtime. The suggested system’s result is to enhance the classifier’s performance in spotting illness early[10].

The investigation into machine learning for diabetes prediction using the PIMA dataset advances the field with key innovations and contributions:

- This research refines diabetes prediction through comprehensive feature engineering that improves data quality, leading to superior predictive accuracy.
- It presents the AdvanSVM, a customized Support Vector Machine, which marks a significant innovation by targeting the specific challenges of medical data within the PIMA dataset.
- The study sets a benchmark by meticulously evaluating the AdvanSVM using key

performance indicators, thereby verifying its effectiveness in the accurate prediction of diabetes.

- By addressing the specific challenges of the PIMA dataset, such as incomplete data and anomalies, the research ensures the AdvanSVM is adept at managing data peculiarities, underpinning its clinical relevance in enhancing diabetes diagnosis and treatment strategies.

The structure of the paper includes sections on related work, methodology, results and discussion, and the conclusion.

Literature work

In this paper [1], four prominent imputation techniques—MICE, KNN, Mean, and Median—and applied them to several classifiers, including Decision Trees, Random Forest, SVC, and Gaussian Naive Bayes, to gauge their effectiveness in diabetes diagnosis within the PIMA dataset. In this research investigation revealed that the choice of imputation method significantly affects the classifiers' predictive performance, with certain techniques aligning better with specific models. Notably, MICE and KNN often outperformed simpler methods like Mean and Median imputation in more complex classifiers. However, this study is limited by the constraints inherent to the dataset itself, such as the underlying distribution of missing data, which may influence the generalizability of our findings[1].

Exploring the use of classification algorithms in machine learning, this article delves into recent implementations on the Indian Survey Database specifically for diabetes detection. Through a comparative analysis based on feature selection and evaluation metrics, it emerges that models integrating feature selection consistently yield higher accuracy[2]. Notwithstanding these advances, the research recognizes limitations regarding dataset specificity and the extrapolation of findings beyond the dataset parameters.

Diabetes mellitus remains a significant global health concern, with approximately half of those affected remaining undiagnosed. Efforts to facilitate early detection and management are crucial for mitigating its complications. Our study leverages advanced machine learning and deep learning techniques, including stacking-based models, to enhance the early diagnosis of diabetes through the analysis of the PIMA Indian diabetes dataset, simulated data, and additional information gathered from a local healthcare facility.

Utilizing both classical and deep neural network stacking ensemble methods, we aim to improve the classification accuracy and robustness by amalgamating predictions from various models. Our findings reveal that a stacking ensemble of three neural network architectures achieves the highest accuracy of 95.50%, with precision, recall, and F1-score rates of 94%, 97%, and 96% respectively, through 5-fold cross-validation on a simulation study. Performance on the PIMA Indian Diabetes dataset shows an accuracy of 75.03% with the train-test split method and 77.10% with cross-validation, indicating a 2.23%–12% improvement using the latter. The proposed methodologies demonstrate promising accuracy rates ranging from 92% to 95%, along with precision, recall, and F1-score ranges of 88% to 96%, using both classical and deep learning stacking approaches on the primary dataset.

However, the study is not without limitations. The reliance on specific datasets, including the simulated data and information from a single healthcare facility, may limit the generalizability of the findings. Additionally, while the accuracy rates are high, further research is needed to validate these models across more diverse and larger populations to ensure their effectiveness in real-world settings[3].

Diabetes Mellitus (DM) is a chronic disease characterized by elevated blood sugar levels, leading to serious health issues such as heart disease, kidney failure, and stroke if unmanaged. Traditional machine learning and deep learning methods have struggled with precise diabetes prediction, particularly with large datasets. To address this, our study introduces a novel classification approach combining Latent Dirichlet Allocation (LDA) and Artificial Neural Networks (ANN), aiming to enhance prediction accuracy through a sophisticated integration of probability distributions and neural network backpropagation.

Using data from the PIMA and North California State University (NCSU) datasets, we applied min-max normalization and bivariate filter-based feature selection to prepare the data. This preprocessing enabled effective feature selection through Pearson correlation, setting the stage for the advanced ANN-LDA classification method. Our approach outperformed existing methods, achieving an impressive 93% classification accuracy[4]. Table 1 describes the comprehensive literature survey summary.

Table 1: Summary of Literature review.

Cite	Methods Used	Dataset Used	Findings	Limitations	Performance Metrics
[1]	MICE,KNN, Mean,Median imputation	PIMA Indian Diabetes dataset	Evaluated impact of imputation methods ondiabetes prediction models	NA	NA

[2]	Machine Learning Classification Algorithms	Indian Survey Database	Feature selection methods improved model accuracy	Limited dataset discussion	More accurate Than other approaches
[3]	Stacking Ensemble Models (Classical and Deep NN)	PIMA Indian Diabetes dataset, Simulated Data, Local Healthcare Data	Stacked ensemble models improved accuracy and robustness in diabetes lassification	NA	Accuracy: 92-95%, Precision:88-96%, Recall:88-96%, F1-Score:88.96%
[4]	ANN-LDA	PIMA Dataset, North California State University (NCSU) dataset	Achieved better classification accuracy using LDA and ANN	Specificity of evaluation criteria and datasets not discussed	Accuracy: 93%
[5]	Optimized SVM, KNN, RF with PSO	Indian Pima diabetes dataset	Hybrid classifiers improved diabetes detection rate	Hyper-parameter tuning complexity	Accuracy: 94.27%
[6]	ANN, RF, GB, Tab-Net, SVM	PIMA and Early-Risk Diabetes datasets	ANN and RF showed remarkable accuracy on PIMA and early-risk diabetes datasets	Limited comparison with existing models	ANN: Accuracy: 99.35% RF:Accuracy:99.36%
[7]	SVM, Decision Tree, RF, KNNetc.	Pima Indians Diabetes Database	Random Forest Showed higher accuracy among used algorithms	Limited discussion on dataset size and feature selection	Accuracy: 91.10%

[8]	BPNN with Batch Normalization	Pima diabetes dataset, CDC BRFSS2015 dataset, Mesra Diabetes dataset	Non-invasive methodshowed significant improvements in diabetes diagnosis	Comparison with traditional method not exhaustive	Accuracy: 89.81-95.28%
[9]	Semisupervised Mode with Gradient Boosting and SMOTE	Pima Indians dataset, Private diabetes datasets	Showed impressive performance in predicting diabetes	Domain adaptation method not detailed	XGBoost with SMOTE: Accuracy: 97.4%, F1: 0.95, AUC: 0.87
[10]	SMOTE and SMO	PIMA Indian Diabetes (PID) dataset	\Improved classifier performance in early disease detection	Lack of discussion on computational efficiency	Correction Rate: 99.07%
[11]	Algorithm for Outlier Removal	Pim Diabetes Dataset	Improved machine learning model accuracy by removing outliers	Comparison with limited techniques	Logistic Regression: Accuracy: 84%, Precision:0.88, Recall: 0.65
[12]	LGBM Feature Selector, Grasshopper Optimization	PIMA diabetes dataset	Enhanced model Accuracy for diabetes prediction	Specific optimization details not provided	NA
[13]	Logistic Regression, Gradient Boosting	Pima dataset, Iraqi society dataset	Highlighted the effectiveness of classification models for diabetes prediction	Limited dataset details	Logistic Regression: Accuracy: 0.77, Gradient Boosting: Accuracy: 0.977
[14]	PSO for Feature Selection	Three medical datasets	Improved accuracy and reduced mistake rates with decision tree, RF, and Naïve Bayes	Limited detail on eature selection criteria	High accuracy with low mistake rate
[15]	KNN Imputed Tri-ensemble Voting Classifier	NA	Showed superior performance over other models with impressive metrics	Limited comparison with models not using KNN imputation	Accuracy: 97.49%, Precision: 98.16%, Recall: 99.35% F1:98.84%
[16]	MCDM Framework	Pima Indian diabetes dataset	Recommended logistic regression for diabetes prediction based on MCDM rankings	NA	NA
[17]	SVM and ANFIS	PIMA Indian diabetes dataset	SVM showed higher accuracy in diabetes detection than ANFIS	Limited to two algorithms comparison	SVM: Accuracy: 85.06%, Specificity: 9.57%, Sensitivity:79.60%
[18]	Random Forest	Pima Indian Diabetes dataset	Indicated potential for optimization in early detection of diabetes	Suggestion for further research on feature selection and model complexity	Accuracy: 87%
[19]	Ensemble of Classifiers	NA	Demonstrated feasibility of using ML models for diabetes prediction	Suggestion for mor extensive datasets and advanced techniques	Accuracy: 85%

[20]	Nature-Inspired Metaheuristic Algorithms	NA	Voting classifier with Smote and Bat Algorithm showed high accuracy	Focused on algorithmic performance without discussing data specifics	Accuracy: 98%
[21]	Logistic Regression, SVM	Dataset from GitHub	Compared two classification algorithms for early diabetes detection	Limited to GitHub data analysis	NA

[22]	ANN, CATBoost, XGBoost, Light GBM	NHANES dataset	XGB-h model outperformed others in predicting diabetes	Focus on prediction without addressing intervention strategies	NA
[23]	CNN-LSTM	Comprehensive dataset for diabetes patients	Achieved impressive accuracy, utperforming traditional models	Limited discussion on model integration challenges	Accuracy: 97%
[24]	Hybrid Model (ANN, AdaBoost, RF,Logistic Regression)	NA	Hybrid model outperformed conventional models in diabetes prediction	NA	Binary Classification: Accuracy: 97-99%, Validation Dataset: 79-89%
[25]	En-RfRsK Ensemble Approach	PIMA diabetes dataset	Outperformed existing ML diabetes prediction systems	Limited discussion on the ensemble model complexity	Accuracy: 88.89%

Materials and Methodology

Dataset

The PIMA Indians Diabetes Dataset focuses specifically on data completeness and the occurrence of anomalies across various clinical indicators. The dataset indicates a significant number of missing entries in several categories, notably 'Skin Thickness' and 'Insulin,' which have high incidences of non-reported values, with 227 and 374 missing entries, respectively. The lack of data may signal challenges in the collection process or may indicate the inherent difficulties in obtaining precise measurements for these variables. In contrast, the 'Diabetes Pedigree Function' and 'Age' categories show complete data records with no missing values, suggesting more reliable methods of collection or greater availability of these measurements.

A closer examination of the dataset reveals a substantial number of outliers in specific medical attributes. For instance, the 'Insulin' category suffers from numerous missing values and also presents 20 outliers, which may be due to a range of factors, from individual physiological variations to potential errors in data recording. The 'Blood Pressure' and 'BMI' categories also have a significant number of outliers, with 8 and 5 respectively. Such outliers are critical to consider in any analytical model, as they can substantially distort the results and impact the accuracy of diabetes predictions. The discovery of these data irregularities highlights the necessity of thorough preprocessing to ensure the robustness of the dataset for analysis. The PIMA Indians Diabetes Dataset is crucial for epidemiological studies focusing on this particular demographic. It exposes significant data collection gaps, especially in the 'Skin Thickness' and 'Insulin' measurements, posing challenges for researchers aiming to use this data for predictive research on diabetes. The high rate of incomplete records, particularly in vital diagnostic and management variables, accentuates the urgent need for methodical data imputation techniques to secure analytical accuracy. Moreover, the presence of outliers in critical metrics such as 'Insulin' and 'Blood Pressure' emphasizes the need for stringent outlier management protocols to ensure that statistical assessments of diabetes prevalence in the Pima Indian community remain reliable and credible.

Table 2: PIMA Diabetes dataset features

Feature	Description	Data Type
Pregnancies	Number of times pregnant	Integer
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Integer
BloodPressure	Diastolic blood pressure (mm Hg)	Integer
SkinThickness	Triceps skinfold thickness (mm)	Integer
Insulin	2-Hour serum insulin (mu U/ml)	Integer
BMI	Body mass index (weight in kg/(height in m)^2)	Float
DiabetesPedigreeFunction	Diabetes pedigree function	Float

Table 2 describes the summary of the PIMA Diabetes dataset features, which are key indicators for diabetes analysis. The table outlines seven distinct medical features: 'Pregnancies,' 'Glucose,' 'Blood Pressure,' 'Skin Thickness,' 'Insulin,' 'BMI,' and 'Diabetes Pedigree Function,' each accompanied by a description and data type. The 'Pregnancies,' 'Glucose,' 'Blood Pressure,' 'Skin Thickness,' and 'Insulin' variables are all recorded as integer data types, indicating discrete measurement values. In contrast, 'BMI' and 'Diabetes Pedigree Function' are represented by floating-point numbers, reflecting their continuous nature.

Preprocessing:

Data preprocessing is a crucial step that involves refining the dataset by removing noisy, unstructured, and erroneous data points. Prior to applying any machine learning algorithm, it is essential to address missing values, detect and handle outliers, apply scaling, and ensure that the dataset is balanced to maintain high data integrity [26-29]. In our analysis, we use the z-score method to quantify the deviation of data points from the average. Further, we apply the interquartile range (IQR) method to identify any outliers that might skew the predictions. We visually validate these outliers using box plots and subsequently adjust them to the median value for consistency. For missing entries, we employ median imputation to foster a more stable analytical environment and mitigate the influence of anomalous values. To address any imbalance in the dataset, we incorporate the Synthetic Minority Over-sampling Technique (SMOTE).

Workflow of SVM-Based Diabetes Prediction Model:

The block diagram presents a streamlined process for predicting diabetes using the Indian PIMA dataset within a machine-learning framework. The initial stage involves input data, specifically the Indian PIMA dataset, which is then subjected to data preprocessing to enhance its quality for model training. Preprocessing might include normalization, handling of missing values, and encoding categorical variables among other techniques.

Following preprocessing, the dataset undergoes a split according to predefined ratios: 70% is allocated for training the model, 15% for validation to fine-tune model parameters, and the remaining 15% is reserved for testing the model's predictive power. The model in question is the Proposed AdvanSVM Model, which likely refers to an advanced or modified Support Vector Machine algorithm tailored for this particular application.

The output of the model is a dichotomous classification of 'Diabetes' or 'Non-Diabetes', representing the prediction results. These results are evaluated using appropriate performance metrics, which assess the effectiveness of the model in accurately predicting diabetes status based on the data provided.

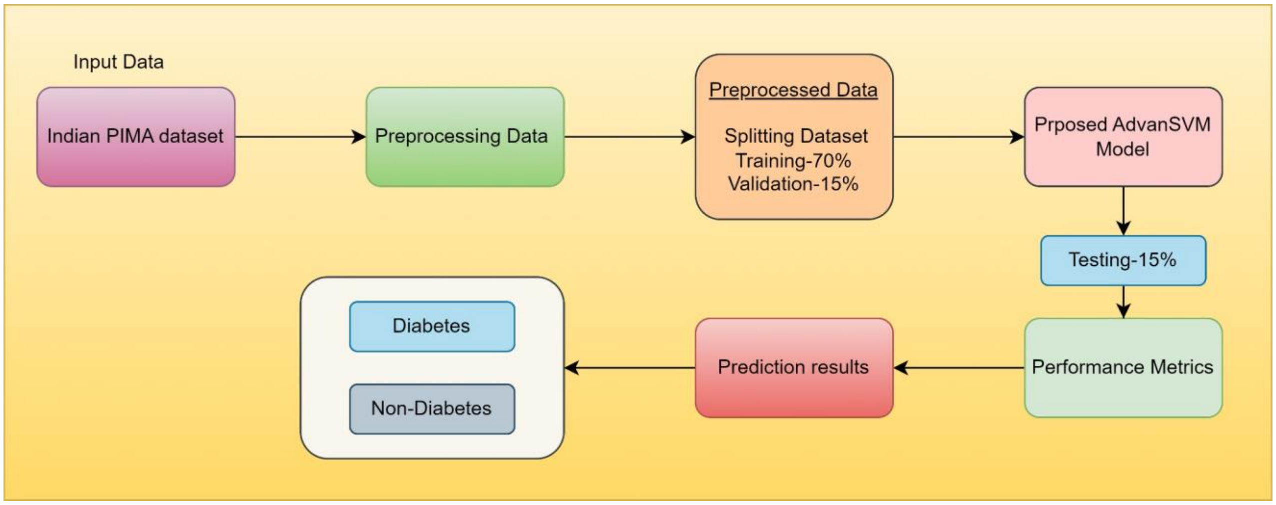

Figure 1:Workflow for Diabetes Prediction Using AdvanSVM Algorithm

Algorithm: Diabetes Prediction using AdvanSVM

```
Input: Indian PIMA dataset, D
Output: Prediction results (Diabetes or Non-Diabetes), Performance Metrics (Accuracy, Precision, Recall, F1-Score)

1: Begin

2: Load the Indian PIMA dataset, D.

3: Preprocess the Data, D':
   a. Impute missing values, if any, in D to obtain D' with no missing values.
   b.Normalize the features in D' to have zero mean and unit variance:
      For each feature column F in D', $F' = (F - mean(F)) / std(F)$
   c. Optionally, perform feature selection to reduce dimensionality to $F' \in R^n$

4: Split the Preprocessed Data, D':
   a. Partition D' into three subsets: D_train, D_val, and D_test.
   b. Assign 70% of D' to D_train.
   b. Assign 15% of D' to D_val.
   c. Assign the remaining 15% of D' to D_test.

5: Develop the Proposed AdvanSVM Model:
   a. Define the SVM objective function to minimize:
      $\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum \xi_i$
      subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$
      where:
```

```
      w is the weight vector,
      b is the bias term,
      ξ_i are the slack variables,
      C is the regularization parameter,
      (x_i, y_i) are the data points with their labels.
   b. Use grid search with K-fold cross-validation on D_train to find the optimal C and kernel parameters (like γ in RBF).
  c. Train the AdvanSVM model on D_train using the optimal parameters found.

6: Evaluate the Model on D_val to adjust parameters if necessary.

7: Test the Model:
  a. Use the model trained with optimal parameters to predict outcomes on D_test.
  b.Calculate Performance Metrics on D_test:
    Accuracy = (TP + TN) / (TP + TN + FP + FN)
    Precision = TP / (TP + FP)
    Recall = TP / (TP + FN)
    F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
    where:
      TP = True Positives, TN = True Negatives,
      FP = False Positives, FN = False Negatives

8: Report the Prediction Results and Performance Metrics.

9: End
```

Results and discussions:

SVM classification:

The AdvanSVM classifier, as suggested by its name, is an advanced iteration of the Advanced Support Vector Machine (AdvanSVM) algorithm, tailored to handle specific characteristics of the Indian PIMA dataset. SVM is a popular machine learning model that is used for classification tasks; it works by finding the hyperplane that best separates the classes in the feature space. The AdvanSVM classifier would incorporate enhancements over the traditional SVM to improve its performance on the PIMA dataset, which may include specialized kernel functions, parameter optimizations, or methods to handle the imbalanced nature of the dataset typically observed in medical data.

In the context of the Indian PIMA dataset, which contains medical measurements for predicting the onset of diabetes, the AdvanSVM classifier would be specifically adjusted to deal with the intricate patterns and distributions inherent in this type of data. This could involve adjusting the SVM to better handle the non-linear relationships by employing more sophisticated kernel functions or integrating techniques like SMOTE (Synthetic Minority Over-sampling Technique) to counteract any imbalance between the diabetic and non-diabetic classes.

Furthermore, since the Indian PIMA dataset has missing values and outliers as indicated in previous discussions, the AdvanSVM classifier may also integrate preprocessing steps that are optimized for this dataset, such as intelligent imputation strategies and robust methods for outlier detection. This ensures that the classifier is not only technically advanced in terms of its algorithmic structure but also fine-tuned to the specific challenges posed by the PIMA dataset, ultimately aiming to yield more accurate and clinically relevant predictions for diabetes onset.

Table 3 : Parameter Tuning Ranges for AdvanSVM on PIMA Dataset

Hyperparameter	**Description**	**Considerations for PIMA Dataset**	**Typical Starting Values or Ranges**
Kernel Type	Type of function to map data into a higher-dimensional space	RBF, polynomial, or sigmoid could be explored	'rbf', 'poly', 'sigmoid', 'linear'
C	Penalty parameter of the error term	To balance classification margin and misclassification rate	0.1, 1, 10, 100
Gamma	Kernel coefficient for 'rbf', 'poly', and 'sigmoid'	To control the trade-off between bias and variance in predictions	scale, 0.1, 0.01, 0.001
Degree	Degree of the polynomial kernel function	Relevant if polynomial kernel is used	2, 3, 4, 5
Coef0	Independent term in kernel function	Adjusts the model's sensitivity to higher-order features	0, 0.5, 1, 10

Class Weights	Weights associated with classes	May be used to address data imbalance	'balanced', or a custom dict
Epsilon	Epsilon in the epsilon-SVR model	Determines the width of the epsilon-tube	0.1, 0.2, 0.5, 1

The Table 3 provides an overview of key hyperparameters for configuring an advanced Support Vector Machine (AdvanSVM) classifier, alongside their descriptions, considerations for the Indian PIMA dataset, and typical starting values or ranges. The kernel type, an essential hyperparameter, determines the transformation function and includes options such as RBF, polynomial, sigmoid, and linear. The penalty parameter, C, influences the trade-off between the classifier's margin and its error tolerance, with a range of values from 0.1 to 100 typically explored. Gamma affects the reach of the RBF, polynomial, and sigmoid kernels, with smaller values denoting a farther reach, and typical starting values ranging from scale to 0.001. Degree is relevant if a polynomial kernel is used, usually varying between 2 and 5. Coef0 is the independent term that can adjust sensitivity to higher-order features in non-linear kernels, and its initial values can be 0, 0.5, 1, or 10. Class weights are utilized to give different importance to classes, particularly useful in imbalanced datasets, with options like 'balanced' or a custom dictionary. Lastly, Epsilon pertains to the epsilon-SVR model and helps to establish the width of the epsilon-tube, with possible starting points being 0.1, 0.2, or 0.5. This tabulation is designed to aid in the systematic tuning of an SVM classifier to optimize its performance on diabetes prediction using the PIMA dataset.

Figure 2 shows a confusion matrix, a tool used to assess the performance of a classification model. The matrix indicates that the model correctly predicted 'Non-diabetic' 81 times and 'Diabetic' 36 times, while incorrectly predicting 18 non-diabetic cases as diabetic and 19 diabetic cases as non-diabetic.

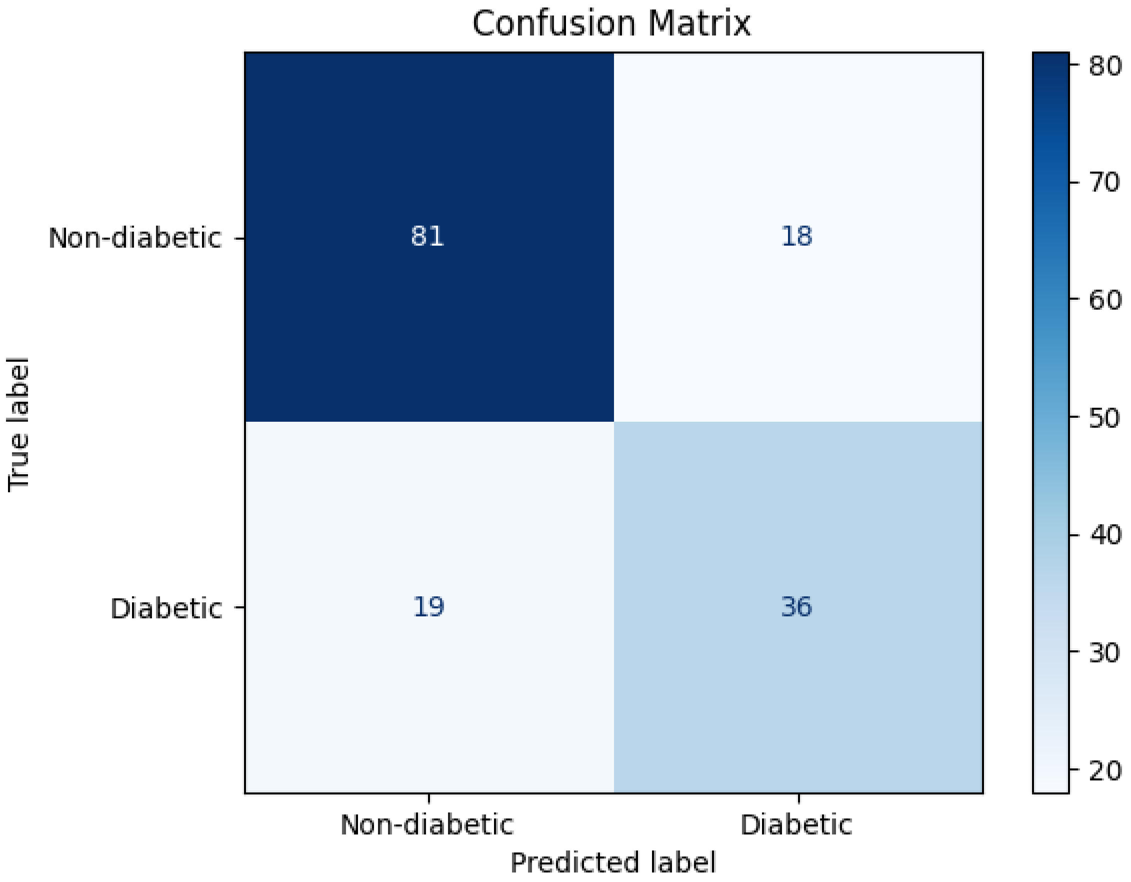

Figure 2: Confusion Matrix

Table 4 summarizing Proposed model AdvanSVM Classification and performance metrics, the precision, recall, f1-score for each class, and the overall accuracy:

Table 4:Performance Metrics

Metric	Class 0 (Non-diabetic)	Class 1 (Diabetic)	Overall
Precision	0.81	0.67	-
Recall	0.82	0.65	-
F1-Score	0.81	0.66	-
Accuracy	-	-	0.76

Conclusion and Future Scope

The comprehensive analysis across various studies demonstrates significant advancements in diabetes prediction utilizing machine learning and data mining techniques. The effectiveness of data imputation methods, and the importance of precise feature selection, underscore the critical role of data preprocessing in enhancing model accuracy. The innovation in combining probabilistic models with neural networks and the utilization of hybrid classifiers and optimization techniques, showcase the evolving landscape of predictive analytics in healthcare. The exploration of non-invasive diagnostic models and semi-supervised learning approaches opens new avenues for early and accessible diabetes detection. Furthermore, the implementation of advanced classification methods and the strategic use of data preprocessing techniques have shown to significantly improve the predictive performance of diabetes models.

The future scope of research in diabetes prediction is vast. There is a growing need to integrate more comprehensive and diverse datasets, including real-world patient data, to further validate and enhance the generalizability of predictive models. The exploration of emerging machine learning algorithms and the development of more sophisticated ensemble methods hold promise for achieving even higher levels of accuracy and reliability. Additionally, the application of explainable AI techniques could provide deeper insights into model predictions, fostering trust and understanding among healthcare providers. Lastly, the potential for deploying these predictive models in mobile applications and healthcare systems underscores the importance of interdisciplinary collaboration to bring these technological advancements to clinical practice, ultimately aiming to improve patient outcomes in diabetes care.

COMPETING INTERESTS

The authors have no compting interest to declare.

Author’s Affiliation

Asha V [1], Manjunath Ramanna Lamani[2],Padmaja K[3] , Tejashwini A Gondhale[4]

[1]Research Scholar, RNS Institute of Technology, Bengaluru
[2]Research Scholar, CHRIST (Deemed to be University), Bengaluru
[3]Assistant Professor, GSSS, Mysore
[4]Assistant Professor, Computer Science Department, Dayananda Sagar College of Engineering, Bengaluru, Karnataka.

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/. *Seybold Report* is a peer-reviewed journal published by Seybold Publications.

HOW TO CITE THIS ARTICLE:

Asha, V., Lamani, M. R., Padmaja, K., & Gondhale, T. A. (2024). Enhancing diabetes prediction accuracy with AdvanSVM: A machine learning approach using the PIMA dataset. *Seybold Report Journal*, *19*(05), 51-72. DOI: 10.5110/77. 1412

REFERENCES

[1]. Jain, Vishesh, Sanyam Shukla, and Nilay Khare. "Analysis of various data imputation techniques for diabetes classification on PIMA dataset." *2024 IEEE International Students' Conference on Electrical, Electronics and ComputerScience (SCEECS)*. IEEE, 2024. [2].Heydari, Farzad, M. Kuchaki Rafsanjani, and M. SHEIKH HOSSEINI LORI. "AN OVERVIEW OF DIABETES DIAGNOSIS METHODS ON THE PIMA INDIAN DATASET."
Journal of Mahani Mathematical Research Center 13.1 (2024).
[3]. Reza, Md Shamim, et al. "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data." *Heliyon* 10.2 (2024).
[4]. Soumya, K. N., and Raja Praveen KN. "Diabetes Mellitus Disease Prediction and Classification using Latent Dirichlet Allocation and Artificial Neural Network Classifier." *International Journal of Intelligent Systems and Applications in Engineering* 12.10s (2024): 98-106.

[5]. Shimpi, Jayanta Kiran, Poonkuntran Shanmugam, and Albert Alexander Stonier. "Analytical model to predict diabetic patients using an optimized hybrid classifier." *Soft Computing* 28.3 (2024): 1883-1892.
[6]. Khan, Qazi Waqas, et al. "An intelligent diabetes classification and perception framework based on ensemble and deep learning method." *PeerJ Computer Science* 10 (2024): e1914.
[7]. Saranya, Gudluri, and Sagar Dhanraj Pande. "Enhancing Diabetes Prediction with Data Preprocessing and various Machine Learning Algorithms." *EAI Endorsed Transactions on Internet of Things* 10 (2024).
[8]. Zhang, Zeyu, et al. "A Deep Learning Approach to Diabetes Diagnosis." *arXiv preprint arXiv:2403.07483* (2024).
[9]. El-Sofany, Hosam, et al. "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App." *International Journal of Intelligent Systems* 2024 (2024).
[10]. Talari, Praveen, et al. "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2." *Plos one* 19.1 (2024): e0292100.
[11].Joglekar, Pushkar, et al. "Skewness based Outlier Elimination Algorithm for Pima Diabetes Dataset." *Grenze International Journal of Engineering & Technology (GIJET)* 10 (2024). [12].Malarvizhi, K., et al. "Prediction of Diabetes Mellitus using Ensemble Methods in Machine Learning." *Grenze International Journal of Engineering & Technology (GIJET)* 10.1 (2024). [13]. Vardhan, Harsh, and Vijay Kumar. "Comparative Analysis of Machine Learning-Based Diabetes Prediction Approaches." *Future of AI in Medical Imaging*. IGI Global, 2024. 65-75.) [14]. Abdollahi, Jafar, and Solmaz Aref. "Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms." *SN Computer Science* 5.2 (2024): 217.

[15]. Alnowaiser, Khaled. "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model." *IEEE Access* (2024).)

[16]. Kumar, Ajay, and Kamaldeep Kaur. "A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction." *International Journal of Engineering & Technology Innovation* 14.1 (2024).

[17]. Sai, A., and R. Puviarasi. "Diabetes mellitus (DM) detection using SVM algorithm and adaptive neuro fuzzy inference system (ANFIS) for accuracy, specificity, and sensitivity improvement." *AIP Conference Proceedings*. Vol. 2816. No. 1. AIP Publishing, 2024.

[18]. Noviyanti, Cindy Nabila, and Alamsyah Alamsyah. "Early Detection of Diabetes Using Random Forest Algorithm." *Journal of Information System Exploration and Research* 2.1 (2024 [19]. . Yadav, Ayush, and Bhuvaneswari Amma NG. "A Smart Healthcare Diabetes Prediction System Using Ensemble of Classifiers." *Using Traditional Design Methods to Enhance AI-Driven Decision Making*. IGI Global, 2024. 118-133.

[20]. Jain, Anjali, and Alka Singhal. "Bio-inspired Approach for Early Diabetes Prediction and Diet Recommendation." *SN Computer Science* 5.1 (2024): 182.

[21]. Sreekumar, et al. "Diabetes Prediction: A Comparison Between Generalized Linear Model and Machine Learning." *Computational Intelligence in Healthcare Informatics*. Singapore: Springer Nature Singapore, 2024. 57-73.

[22]. Sarkar, Prosanjeet Jyotirmay, et al. "Prediction model for diabetes mellitus using machine learning algorithms for enhanced diagnosis and prognosis in healthcare." *Computer and Telecommunication Engineering* 2.1 (2024): 2446.

[23]. Ayat, Yassine, et al. "Novel diabetes classification approach based on CNN-LSTM: enhanced performance and accuracy." *Diagnostyka* 25 (2024).

[24]. Zohair, Mohammad, et al. "A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification." *International Journal of Information Technology* 16.3 (2024): 1955-1965.

[25]. NG, Bhuvaneswari Amma. "En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus." *Egyptian Informatics Journal* 25 (2024): 100441.

[26] R. Amin, R. Yasmin, S. Ruhi, H. Rahman, S. Reza, Informatics in Medicine Unlocked Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms, Inform. Med. Unlocked 36 (June 2022) (2023) 101155, https://doi.org/10.1016/j.Imu.2022.101155 .

[27] R. Yasmin, R. Amin, M.S. Reza, Effects of hybrid non-linear feature extraction method on different data sampling techniques for liver disease prediction, J. Futur. Sustain. 2 (2) (2022) 57–64, https://doi.org/10.5267/j.jfs.2022.9.005 .

[28] X.H. Cao, I. Stojkovic, Z. Obradovic, Open Access A robust data scaling algorithm to

improve classification accuracies in biomedical data, BMC Bioinf. (2016) 1–10, https://doi.org/10.1186/s12859-016-1236-x .

[29] S. Fotouhi, S. Asadi, M.W. Kattan, A comprehensive data level analysis for cancer diagnosis on imbalanced data, J. Biomed. Inf. 90 (December 2018) (2019) 103089, https://doi.org/10.1016/j.jbi.2018.12.003 .