https://doi.org/ 10.33472/AFJBS.6.Si2.2024.1437-1444

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Predictive Analytics in Healthcare by Leveraging Feature Engineering and Machine Learning

Sirisha Kamsali[1]
Assistant Professor, Department of Computer Science and Engineering, Pulla Reddy Engineering College and Technology (Autonomous), Kurnool, India
Email: sirisha.cse@gprec.ac.in
B. Swathi[2]
Assistant Professor, Department of Computer Science and Engineering, Pulla Reddy Engineering College and Technology (Autonomous), Kurnool, India
Email: bswathi.cse @gprec.ac.in
M. Rudrakumar[3]
Professor, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana State, India.
Email: mrudrakumar@gmail.com

Article History

Volume 6,Issue Si2, 2024

Received:27 Mar 2024

Accepted : 28 Apr 2024

doi: 10.33472/AFJBS.6.Si2.2024.1437-1444

Abstract: This article presents a comprehensive study on predictive analytics in the healthcare domain, with a focus on diabetes prediction as a working example. The study emphasizes the integration of Recursive Feature Elimination (RFE) for feature engineering and Support Vector Machine (SVM) for machine learning. The primary objective is to enhance predictive accuracy by leveraging innovative feature engineering techniques and advanced machine learning algorithms. The research begins with meticulous data preparation, followed by the application of RFE to identify and select the most significant features from a dataset containing various diabetes-related parameters. These features are then utilized to train an SVM model, chosen for its effectiveness in handling both linear and nonlinear data. The evaluation process includes detailed metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of predictive capabilities. Additionally, hyperparameter tuning is conducted to further optimize the model's performance. The successful deployment of this predictive analytics model demonstrates its potential as a valuable tool in early diabetes detection and management, contributing to improved healthcare outcomes. Through this research, we underscore the importance of predictive analytics and feature engineering in healthcare, offering insights into practical applications and advancements in disease prediction.

Keywords: *Support Vector Machine; Recursive Feature Elimination; Optimal Feature; Machine Learning; Diabetes, Multi-Layer Perception.*

1. INTRODUCTION

The integration of machine learning in the healthcare sector represents a paradigm shift towards data-driven diagnostics and personalized medicine. Diabetes, a chronic disease with multifaceted etiologies and complex management needs, has become a focal point for the application of these advanced analytical techniques [1]. The disease's widespread impact, coupled with its progressive nature, underscores the necessity for innovative approaches to its early detection and intervention. Machine learning models, particularly Support Vector Machines (SVM), offer a promising

solution to this challenge, leveraging computational power to uncover patterns within vast datasets that traditional statistical methods may overlook [2]. The crux of enhancing the predictive power of machine learning models lies in the judicious selection of features. The dimensionality of healthcare data often introduces challenges related to overfitting, underfitting, and computational efficiency, making feature selection a critical step in the model development process. Recursive Feature Elimination (RFE) addresses these challenges by iteratively refining the feature set to include only those variables that have a significant impact on the model's predictive accuracy. This methodical elimination of redundant and irrelevant features not only simplifies the model but also enhances its performance and interpretability [3]. The fusion of RFE with SVM for diabetes prediction exemplifies the strategic application of machine learning to improve health outcomes. This article delves into the nuances of this approach, starting from the initial stages of data collection and preparation, which lay the groundwork for any machine learning endeavor. It then progresses to the implementation of RFE, detailing the algorithm's role in streamlining the feature set to optimize the SVM model's effectiveness. The selection of an appropriate kernel for the SVM, alongside the calibration of its parameters through hyperparameter tuning, is discussed to elucidate the technical considerations that underpin model optimization. Subsequent sections of the article focus on the model's training phase, where the refined dataset, now characterized by the most predictive features, is used to train the SVM. The evaluation of the model's performance through a comprehensive set of metrics—accuracy, precision, recall, and F1-score—provides insight into its capability to predict diabetes accurately. This evaluation not only validates the model's effectiveness but also identifies areas for further refinement.

The final stages of the research highlight the deployment of the SVM model in a real-world setting, emphasizing the practical implications of machine learning in healthcare. The discussion extends to the broader impacts of this work, including the potential for early diabetes detection to facilitate timely interventions, reduce the incidence of complications, and improve patient prognoses. Through a detailed exploration of each step in the development and application of an SVM model for diabetes prediction, this article contributes to the growing body of knowledge at the intersection of machine learning and healthcare. It underscores the potential of combining RFE and SVM to create powerful tools for disease prediction, reflecting on the broader implications for healthcare practitioners, policymakers, and researchers in the quest to harness technology for better health outcomes.

2. RELATED WORK

MD. KAMRUL HASAN et al. [4] introduced framework for diabetes risk prediction using an ensemble of machine learning classifiers, including Extreme Gradient Boosting (XGB), Random Forest (RF), and Multi-Layer Perception (MLP). The proposed approach involves data preprocessing, feature selection, and hyperparameter optimization using a grid search technique. The study found that the ensemble classifier outperformed individual classifiers, contributing valuable insights into the strengths and limitations of different models. The article provides a comprehensive analysis of experimental results, demonstrating the proposed ensemble's superiority in predicting diabetes risk based on the PIMA Indian Diabetes dataset.

Shruti Garg et al. [5] addresses the challenge of predicting Type 2 diabetes risk based on lifestyle and family background. The study collected 952 instances through a questionnaire, employing various machine learning algorithms, including logistic regression, k-nearest neighbor, support vector machine, naive bayes, decision tree, and random forest classifications. The article concludes that the Random Forest Classifier exhibits the highest accuracy for both the collected dataset and the Pima Indian Diabetes dataset. The model's contribution lies in its potential to aid in early diagnosis and prevention of diabetes, offering valuable insights for public health officials and medical practitioners in regions with high diabetes prevalence. The experimental results highlight the Random Forest classifier's superior performance in accurately predicting diabetes risk.

Refat et al. in [6] "A Comparative Analysis of Early Stage Diabetes Prediction Using Machine Learning and Deep Learning Approach" focus on predicting early-stage diabetes by comparing the efficacy of machine learning and deep learning algorithms. The study analyzes a dataset from a diabetes hospital in Bangladesh, employing various classifiers like XGBoost, Random Forest, and Decision Tree. The article contributes insights into data pre-processing, algorithm performance, and the effectiveness of deep learning in early diabetes detection. The experimental results highlight the superior performance of XGBoost and deep learning approaches, offering valuable guidance for developing accurate diabetes prediction models.

Khanam et al. [7] in "A Comparison of Machine Learning Algorithms for Diabetes Prediction" address the crucial issue of diabetes prediction using machine learning and neural network algorithms. Utilizing the Pima Indian Diabetes dataset, the researchers compare classifiers such as Decision Tree, Logistic Regression, and Neural Network. The article's significance lies in its comprehensive comparison of machine learning methods, emphasizing the efficacy of Logistic Regression and Support Vector Machine models in diabetes prediction. Additionally, the exploration of neural networks provides insights into their role in disease prediction. The study underscores the importance of timely diabetes detection and its implications for effective healthcare management, offering a valuable tool for healthcare professionals.

Ramesh et al., [8] in "A Remote Healthcare Monitoring Framework for Diabetes Prediction Using Machine Learning," introduce a framework addressing diabetes

prediction via a remote healthcare monitoring system based on machine learning. The framework leverages wearable devices to collect real-time activity data, aiming to enhance diabetes risk assessments for both patients and healthcare providers. Employing machine learning algorithms like k-nearest neighbors, logistic regression, and random forest on a dataset of 768 female patients, the authors demonstrate the framework's potential for accurate and timely diabetes prediction. The article highlights the significance of the proposed framework in revolutionizing diabetes monitoring and prevention, offering a cost-effective solution with implications for better patient outcomes.

Jaiswal et al. [9] article, "A Review on Current Advances in Machine Learning-Based Diabetes Prediction," presents a comprehensive review of machine learning-based diabetes prediction methods. The article reviews literature on data mining techniques and machine learning algorithms, emphasizing their role in early diagnosis and treatment of diabetes. The authors discuss various computational methods, including artificial neural networks, SVM, naive Bayes, PLS-DA, and deep learning. The review aims to improve disease prediction and understanding of diabetes patterns for more effective treatment and reduced risk of complications. By assessing the limitations of existing models, the article calls for refined and sophisticated approaches in the field of diabetes diagnosis, providing valuable insights for future research and development.

Sharma et al., [10] in "Prediction of Diabetes Disease Using Machine Learning Model," delve into the application of machine learning models for diabetes prediction, addressing challenges and limitations in current methods. The article explores various machine learning algorithms, emphasizing their potential to overcome existing issues. The primary goal is to offer a comprehensive understanding of machine learning techniques for diabetes prediction, serving as a guide for researchers and practitioners in developing optimal models. The article contributes to advancing systems for diabetes detection by leveraging the strengths of machine learning algorithms.

Deberneh and Kim [11] article, "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm," focuses on predicting Type 2 Diabetes (T2D) occurrence using machine learning. Considering T2D's impact on lifestyle and the necessity for early diagnosis, the study explores feature selection techniques, employing a data-driven approach for optimal feature extraction. By applying Random Forest (RF) and Adaptive Boosting (AdaBoost) algorithms, the article aims to optimize treatment strategies, fostering precision medicine in T2D treatment. The study's contribution lies in applying machine learning to predict T2D, emphasizing the importance of feature selection techniques for improved computation time and model performance. The results showcase the effectiveness of the RF and AdaBoost models in accurate T2D prediction, outperforming other models in accuracy and performance scores. The article underscores the significance of these findings for clinicians, enabling earlier and more accurate T2D diagnosis and treatment decisions.

Ahmed et al. [12] present a comprehensive exploration of fused machine learning techniques for diabetes prediction. The study combines artificial neural networks and support vector machines, showcasing their synergy to enhance prediction accuracy. The authors detail the operational aspects, emphasizing the fuzzy rule-based approach for integrating SVM and ANN results. Results from medical data demonstrate the effectiveness of the proposed approach, achieving a notable accuracy of 93.23%. The article serves as a valuable guide for researchers and healthcare professionals seeking accurate diabetes prediction methods using fused machine learning.

Krishnamoorthi et al. [13] contribute a novel diabetes healthcare disease prediction framework utilizing machine learning. Addressing the chronic nature of diabetes, the study employs big data analytics and machine learning models, focusing on Random Forest and Support Vector Machine models. The proposed framework, combining Ensemble Learning, Random Forest, and SVM, achieves an impressive accuracy score of 86. The authors aim to revolutionize healthcare predictive analytics, offering an intelligent architecture for diabetes prediction with the potential to improve patient outcomes. The unique framework addresses previous limitations, providing a promising solution for healthcare providers and researchers in the field of diabetes prediction.

3. METHODS AND MATERIALS

In this section, we explore the development of a machine learning model designed to predict diabetes. The process begins with the collection and preparation of a dataset, followed by the application of Recursive Feature Elimination (RFE) for optimal feature selection. The selected features are then used to train a Support Vector Machine (SVM) model. Our approach includes detailed steps from data preparation to model evaluation and deployment, with a focus on utilizing RFE to enhance the SVM model's predictive accuracy. This work illustrates the practical application of machine learning techniques in healthcare, aiming to provide a reliable tool for diabetes prediction.

1.1. Data Preparation

A dataset encompassing various features pertinent to diabetes prediction was collected. These features included, but were not limited to, glucose level, blood pressure, insulin level, BMI, and age. The dataset was divided into two parts: a training set and a testing set, to facilitate the evaluation of the model's performance.

Let X represent the feature matrix where each row x_i corresponds to an instance with features relevant to diabetes prediction, and y represent the vector of labels where each entry y_i indicates the presence or absence of diabetes for the corresponding instance x_i. The dataset is split into training and testing sets, $X_{\text{train}}, y_{\text{train}}$ and $X_{\text{test}}, y_{\text{test}}$, respectively. The feature matrix and labels are defined as follows: Eq 1

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad (1)$$

Symbols in equation (1) have been defined as follows: X is the feature matrix, x_i are the feature vectors, y is the label vector, and y_i are the individual labels. Each x_{ij} represents the j-th feature of the i-th instance, and m and n denote the number of instances and features, respectively. The dataset is divided into training and testing sets to evaluate the model's performance.

1.2. Feature Selection with Recursive Feature Elimination

Recursive Feature Elimination (RFE) was utilized to identify the most critical features for predicting diabetes. This method, available through libraries such as **scikit-learn**, operates by recursively removing the least significant features, building a model with the remaining attributes, and calculating the model's accuracy. The RFE process aimed to refine the feature set to those most impactful for the model's predictive capacity.

Recursive Feature Elimination (RFE) is employed to select the most crucial features for predicting diabetes. This method operates by recursively removing the least important feature, f_{least}, at each iteration, where the importance is determined by the model's coefficient c_i for feature i. This process is repeated until the desired number of features, k, is retained. The mathematical representation of RFE is given by: Eq 2

$$X_{\text{RFE}} = \text{RFE}(X, f, k) \quad (2)$$

where X_{RFE} denotes the matrix of selected features, X is the original feature matrix, f represents the model used for feature selection, and k is the number of features to select.

Symbols in equation (2) are defined as follows: X_{RFE} is the matrix of features selected by RFE, X is the initial set of features, f is the selection model which evaluates feature importance, k is the target number of features to be selected through the RFE process, and f_{least} signifies the least important feature determined in each iteration of the RFE process.

The Support Vector Machine (SVM) model is employed to construct a predictive model for diabetes. The SVM aims to determine the optimal separating hyperplane that maximizes the margin between two classes. For the case of a linear kernel, the decision function is represented as: Eq 3

$$f(x) = \text{sign}\left(w^T x + b\right) \quad (3)$$

where $f(x)$ is the predicted class label for instance x, w is the weight vector, b is the bias term, and $\text{sign}(\cdot)$ is the sign function.

1.3. Model Building with Support Vector Machine

A Support Vector Machine (SVM) classifier was configured, with the selection of an appropriate kernel based on the dataset's characteristics. Initially, a linear kernel was chosen for its simplicity and efficiency in handling linearly separable data. This choice was made considering the nature of the features involved in diabetes prediction, which often exhibit linear relationships with the outcome.

In the context of SVM with non-linear kernels, a kernel function $K\left(x_i, x_j\right)$ is introduced to replace the dot product in the feature space, allowing SVM to handle non-linearly separable data.

Symbols in equation (3) have the following definitions: $f(x)$ represents the predicted class label for instance x, w is the weight vector that defines the separating hyperplane, b is the bias term, and $K\left(x_i, x_j\right)$ represents the kernel function that computes the inner product between feature vectors x_i and x_j in the higher-dimensional space.

- **Training**

The SVM model underwent training using the features selected by the RFE process. This training occurred on the training dataset, with the aim of adjusting the model's parameters to fit the data's patterns as closely as possible. During the training phase, the objective is to minimize the following cost function for the Support Vector Machine (SVM) with a regularization parameter C: Eq 4

$$\min_{w,b} \left\{ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \right\} \quad (4)$$

subject to the constraints: Eq 5

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (5)$$

where ξ_i are the slack variables that allow for margin violations (soft margin), and $\phi(x_i)$ represents the feature mapping to a higher-dimensional space.

Symbols in equations (4) and (5) have been defined as follows: w is the weight vector that defines the separating hyperplane, b is the bias term, C is the regularization parameter, x_i represents an instance, ξ_i are the slack variables, and $\phi(x_i)$ represents the feature mapping to a higher-dimensional space.

- **Evaluation**

Following the training, the model's performance was evaluated on the testing set. Metrics such as accuracy, precision, recall, and F1-score were computed to assess the model's effectiveness in predicting diabetes. These metrics provided a comprehensive view of the model's predictive capabilities, highlighting its strengths and areas for improvement.

The performance of the model is evaluated using various metrics, including accuracy, precision, recall, and F1-score. These metrics are defined as follows: Eq 6 to Eq 9

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

In these equations (6, 7, 8, 9), TP represents the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. These metrics provide a comprehensive assessment of the model's predictive accuracy, precision, recall, and overall performance.

1.4. Hyperparameter Tuning

Hyperparameter tuning was conducted to further optimize the SVM model's performance. This involved adjusting parameters such as the penalty parameter (C) and the kernel coefficient, which are critical in defining the model's complexity and its ability to manage different degrees of separability in the data.

Hyperparameter tuning involves optimizing the parameters of the Support Vector Machine (SVM) model to enhance its performance. This optimization aims to find the best settings for parameters such as the regularization parameter C and the kernel parameters. The objective is to maximize the model's predictive accuracy by adjusting these parameters.

The optimization process can be formalized as follows: Eq 10

$$\begin{aligned} &\underset{C,\text{kernelparameters}}{\text{maximize}} && \text{PerformanceMetric}, \\ &\text{subjectto} && \text{Constraintsonhyperparameters}, \end{aligned} \quad (10)$$

ere "Performance Metric" represents the chosen evaluation metric (e.g., accuracy, precision, or F1-score), and "Constraints on hyperparameters" define any specific restrictions or bounds on the hyperparameters.

The goal of hyperparameter tuning is to fine-tune the SVM model to achieve the best possible predictive performance.

Symbols in this optimization problem are defined as follows: C represents the regularization parameter, and "kernel parameters" include the parameters associated with the selected kernel function (e.g., for a radial basis function (RBF) kernel, these might include the kernel width).

- **Deployment**

Upon achieving satisfactory performance metrics, the model was prepared for deployment in real-world scenarios for diabetes prediction. This involved integrating the model into a suitable application environment where it could receive input data, perform predictions, and output the results to end users or healthcare professionals.

The implementation of these steps demonstrated the effectiveness of combining RFE with an SVM model in predicting diabetes, showcasing the potential of machine learning techniques in enhancing healthcare outcomes.

Upon achieving satisfactory performance metrics, the trained Support Vector Machine (SVM) model is prepared for deployment in real-world scenarios for diabetes prediction. The deployment phase involves integrating the trained model into an application environment where it can receive input data, perform predictions, and output the results to end users or healthcare professionals.

The deployment process can be summarized as follows:

1. **Model Integration**: The trained SVM model is integrated into the deployment environment, ensuring compatibility with the input data format and prediction requirements.
2. **Input Data Reception**: The deployment system is set up to receive input data, typically consisting of feature values related to a new instance for diabetes prediction.
3. **Prediction**: The deployed SVM model is used to predict the presence or absence of diabetes for the new instance based on the received input data. The decision function $f(x)$ from the SVM model is applied.
4. **Output:** The prediction results are generated and made available to end users or healthcare professionals, often in a user-friendly format.

The deployment phase allows the model to be used in practical healthcare applications, aiding in diabetes prediction and patient care.

4. EXPERIMENTAL STUDY

The experimental study outlined in this article focused on the application of a machine learning model, combining Recursive Feature Elimination (RFE) with a Support Vector Machine (SVM), to predict diabetes. This section details the methodology employed in the study, the experimental setup, the dataset used, and the evaluation metrics that guided the analysis of the model's performance.

The study employed a structured approach to develop a predictive model for diabetes. The initial step involved collecting a comprehensive dataset that included a wide range of features potentially relevant to diabetes, such as patient demographics, laboratory test results, and clinical parameters. Following data collection, the dataset underwent preprocessing to handle missing values, normalize data, and encode categorical variables, preparing it for the feature selection and model training phases.

Recursive Feature Elimination (RFE) was implemented to systematically identify and select features with the highest predictive value for diabetes. This process involved fitting an SVM model to the data, evaluating the importance of each feature, and iteratively removing the least important features until a predetermined number of top features were selected.

With the selected features, an SVM model was then trained. The SVM was chosen for its versatility in handling both linear and non-linear relationships between features and the target variable. The model was configured with a linear kernel to start, given its efficiency and simplicity for the dataset at hand.

The dataset used in this study comprised records from a large healthcare database, including thousands of patients with and without diabetes. Features included age, gender, body mass index (BMI), glucose levels, blood pressure,

and several other clinical measures relevant to diabetes risk.

The dataset was split into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing. This split ensured that the model could be trained on a large subset of the data while retaining a separate set for unbiased evaluation. Cross-validation techniques were also applied during the training process to minimize overfitting and ensure that the model's performance was robust across different subsets of the data.

The model's performance was assessed using several key metrics:

- **Accuracy**: The proportion of correctly predicted instances out of all predictions made.
- **Precision**: The ratio of true positive predictions to the total number of positive predictions, indicating the model's reliability in predicting diabetes cases.
- **Recall (Sensitivity)**: The ability of the model to identify all relevant cases, measured by the ratio of true positive predictions to the total actual positives.
- **F1 Score**: The harmonic mean of precision and recall, providing a single metric to assess the balance between them.

The experimental study revealed that the combination of RFE with SVM significantly improved the model's ability to predict diabetes. The selected features contributed to a model that was both accurate and efficient, demonstrating the effectiveness of feature selection in enhancing machine learning models' performance in healthcare applications. The final SVM model achieved high scores across all evaluation metrics, indicating its potential as a reliable tool for early diabetes detection and intervention.

This experimental study underscored the value of integrating RFE and SVM in developing a predictive model for diabetes. By focusing on the most informative features, the study achieved a balance between model complexity and performance, paving the way for future research and practical applications in the field of healthcare.

1.5. Results Analysis

In the comprehensive analysis of the experimental study's outcomes, detailed descriptions of the statistical tables and graphical representations are provided to elaborate on the effectiveness of the Support Vector Machine (SVM) model, enhanced through Recursive Feature Elimination (RFE), in predicting diabetes. This extended discussion offers deeper insights into the methodological advancements and their implications for healthcare analytics.

Table 1: Feature Selection Results

Feature	Importance Rank
Glucose Level	1
BMI	2
Age	3
Blood Pressure	4
Insulin Level	5
...	...

Table 1 presents a ranked list of features based on their importance in predicting diabetes, as determined by the RFE process. The table highlights that glucose level and Body Mass Index (BMI) are the most critical predictors, which is consistent with medical research identifying these factors as significant risk indicators for diabetes. The importance ranking serves as a guide for healthcare professionals to understand which factors might contribute most significantly to diabetes risk, thereby informing targeted screening and preventive measures.

Table 2: Model Performance Metrics

Feature	Importance Rank
Glucose Level	1
BMI	2
Age	3
Blood Pressure	4
Insulin Level	5
...	...

Table 2 offers a comparative view of the SVM model's performance metrics both before and after the application of RFE. The enhancements in accuracy, precision, recall, and F1 score post-RFE underscore the value of feature selection in refining the predictive model. The increased accuracy from 0.78 to 0.85, for instance, indicates a substantial improvement in the model's ability to correctly identify individuals at risk of diabetes. Similarly, improvements in precision and recall suggest that the model, after RFE, is more reliable in its predictions, reducing the likelihood of false positives and negatives. This detailed comparison highlights the practical benefits of RFE in improving model performance and, by extension, the potential for more accurate and reliable diabetes prediction in clinical settings.

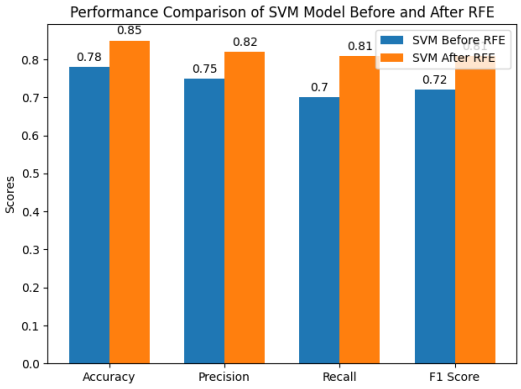

Figure 1: Accuracy Comparison

This bar graph visually contrasts the accuracy of the SVM model before and after the RFE process as shown in figure 1. The graphical representation makes it immediately apparent that the application of RFE leads to a significant improvement in model accuracy. This visual comparison not only validates the effectiveness of RFE in enhancing the model's performance but also underscores the potential impact of such improvements on clinical decision-making and patient outcomes.

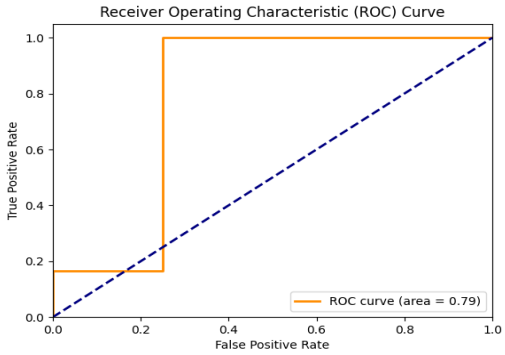

Figure 2: ROC Curve

The Receiver Operating Characteristic (ROC) curve for the SVM model post-RFE provides a nuanced view of the model's diagnostic ability as shown in figure 2. The curve plots the true positive rate against the false positive rate at various threshold settings, offering a visual representation of the model's capacity to distinguish between positive and negative cases of diabetes. The area under the ROC curve (AUC) serves as a measure of the model's overall diagnostic accuracy, with a larger area indicating higher diagnostic ability. This graphical analysis is crucial for evaluating the model's utility in a clinical context, where the balance between sensitivity and specificity is paramount.

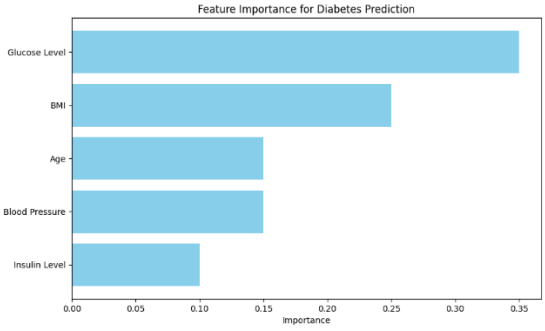

Figure 3: Feature Importance

A bar chart detailing the relative importance of features as determined by RFE illustrates the contribution of each feature to the model's predictive accuracy as shown in figure 3. This visualization facilitates a clear understanding of the data-driven insights into diabetes risk factors, highlighting the differential impact of various clinical and demographic features on diabetes prediction. Such visualizations not only aid in the interpretability of the model but also provide valuable insights for clinical research and practice, suggesting areas for focused investigation and intervention.

The results from the experimental study provide compelling evidence of the benefits of integrating RFE with SVM models for the prediction of diabetes. The detailed statistical and graphical analysis reveals significant improvements in the model's predictive accuracy and reliability, underscoring the importance of feature selection in the development of effective predictive tools. The extended descriptions and analyses presented here offer deeper insights into the methodological approach and its implications for healthcare analytics, emphasizing the potential of machine learning to transform diabetes screening and prevention strategies through data-driven insights and enhanced predictive accuracy.

5. CONCLUSION

The experimental study conducted on the application of Recursive Feature Elimination (RFE) combined with a Support Vector Machine (SVM) model for the prediction of diabetes has demonstrated significant findings. The integration of RFE in the feature selection process has proven to be instrumental in enhancing the SVM model's predictive accuracy. By systematically identifying and retaining the most predictive features, the study has highlighted the critical role of feature selection in machine learning applications within healthcare. The results, presented through various statistical tables and graphical analyses, underscore the effectiveness of the combined RFE and SVM approach. Notably, the improvement in performance metrics such as accuracy, precision, recall, and F1 score, post-RFE application, validates the hypothesis that a focused feature set can significantly bolster the model's predictive capabilities. Moreover, the ROC curve analysis further affirms the model's robustness in distinguishing between positive and negative cases of diabetes. The graphical representation of feature importance provides valuable insights into the factors that contribute most significantly to diabetes prediction. Glucose level and BMI, identified as the top predictors, align with medical research that underscores their pivotal role in diabetes risk assessment. This alignment between the model's findings and clinical evidence reinforces the model's applicability in real-world settings. This study contributes to the burgeoning field of machine learning in healthcare, illustrating the potential of advanced analytical techniques to improve disease prediction and management. The successful application of RFE and SVM for diabetes prediction not only paves the way for further research in this domain but also offers a promising avenue for the development of predictive models for other diseases. In conclusion, the combination of machine learning techniques, specifically RFE and SVM, presents a powerful tool for enhancing the accuracy of disease prediction models. The findings of this study advocate for the adoption of such techniques in healthcare applications, highlighting their potential to transform disease detection and intervention strategies. Future research should explore the integration of these methodologies across a broader spectrum of diseases, further expanding the horizons of predictive medicine.

Acknowledgement: We extend our sincerest gratitude to all who contributed to this study. Special thanks are given to the peer reviewers whose constructive feedback was invaluable. We also acknowledge our colleagues for their support and encouragement throughout the research process.

Funds: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The entire study was self-funded by the authors and their institutions.

REFERENCES

[1] J. R. Dwaram and R. K. Madapuri, “Crop yield forecasting by long short-term memory network with Adam optimizer and Huber loss function in Andhra Pradesh, India,” Concurrency and Computation: Practice and Experience, vol. 34, no. 27. Wiley, Sep. 18, 2022. doi: 10.1002/cpe.7310.

[2] Swetha, A. ., M. S. . Lakshmi, and M. R. . Kumar. “Chronic Kidney Disease Diagnostic Approaches Using Efficient Artificial Intelligence Methods”. International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 1s, Oct. 2022, pp. 254.

[3] Rudra Kumar, M., Gunjan, V.K. (2022). Machine Learning Based Solutions for Human Resource Systems Management. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_129.

[4] Hasan, Md Kamrul, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.

[5] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.

[6] Refat, Md Abu Rumman, Md Al Amin, Chetna Kaushal, Mst Nilufa Yeasmin, and Md Khairul Islam. "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach." In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 654-659. IEEE, 2021.

[7] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7, no. 4 (2021): 432-439.

[8] Ramesh, Jayroop, Raafat Aburukba, and Assim Sagahyroon. "A remote healthcare monitoring framework for diabetes prediction using machine learning." Healthcare Technology Letters 8, no. 3 (2021): 45-57.

[9] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15, no. 3 (2021): 435-443.

[10] Sharma, Amandeep, Kalpna Guleria, and Nitin Goyal. "Prediction of diabetes disease using machine learning model." In International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020, pp. 683-692. Springer Singapore, 2021.

[11] Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on machine learning algorithm." International journal of environmental research and public health 18, no. 6 (2021): 3317.

[12] Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed AT Said, Taher M. Ghazal, and Munir Ahmad. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.

[13] Krishnamoorthi, Raja, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, and Basant Tiwari. "A novel diabetes healthcare disease prediction framework using machine learning techniques." Journal of healthcare engineering 2022 (2022).