https://doi.org/ 10.33472/AFJBS.6.Si2.2024.2638-2649

African Journal of Biological Sciences

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

Research Paper Open Access

A Composite Machine Learning Approach to Predict Diabetes

D Jayanarayana Reddy[1]
Assistant Professor, Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: djnreddy@gmail.com

S. K Jahnavi[2]
Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: jahnavisk2004@gmail.com

S Anudeepthi[3]
Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: seelamanudeepti@gmail.com

Patenge Kanishka[4]
Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: p.g.s.kanishka@gmail.com

K Jithendra[5]
Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: jithendrajeeth3983@gmail.com

M Anis[6]
Department of Computer Science and Engineering, G Pullaiah College of Engineering and Technology (Autonomous), Kurnool, India
Email: anismeerja@gmail.com

·

Article History
Volume 6,Issue Si2, 2024
Received:26 Mar 2024
Accepted : 02 May 2024
doi: 10.33472/AFJBS.6.Si2.2024.2638-2649

Abstract: This article presents a composite machine learning model designed to predict diabetes by leveraging multiple predictive algorithms. The objective of this research is to develop a reliable predictive tool that can be utilized in healthcare settings to facilitate early diagnosis and improved management of diabetes. The model integrates several machine learning techniques to handle diverse data characteristics and improve prediction accuracy. We initiated our approach by gathering a broad dataset from various healthcare databases to ensure a comprehensive demographic and biological representation. The data underwent rigorous preprocessing to ensure consistency and relevance for model training. Feature selection was systematically performed to identify the most significant predictors of diabetes, focusing on reducing model complexity while maintaining predictive integrity. The core of our model architecture combines several machine learning techniques, each selected to complement the others' strengths and weaknesses. This composite approach allows for robustness against overfitting and enhances the generalizability of the model across different patient populations. The performance of the model was evaluated using standard metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). Results from the testing phase demonstrate that our model achieves superior performance in diabetes prediction compared to traditional single-algorithm approaches. This study contributes to the field by providing a scalable and effective framework for diabetes prediction, which can be adapted for further use in other chronic disease contexts. Future work will focus on integrating real-time data analysis and exploring the impact of longitudinal patient data on prediction accuracy.
Keywords: *Machine Learning, Predict Diabetes, Ensemble Model, Support Vector Machine, Recursive Feature Elimination*

1 Introduction

Diabetes mellitus remains one of the most challenging public health issues globally, affecting millions of individuals with its chronic complications that can lead to severe outcomes such as cardiovascular diseases, renal failure, and retinal damage. Early detection and accurate prediction of diabetes can significantly alter the management and intervention strategies, potentially reducing the incidence of such severe complications. The advent of machine learning (ML) in healthcare presents a transformative opportunity to enhance the prediction and diagnostic processes of diabetes.

Recent studies have shown that machine learning techniques can effectively predict diabetes by analyzing various health indicators and patterns from large datasets. Choubey et al. [1] demonstrated the efficacy of using machine learning models to predict diabetes, achieving significant accuracy and providing insights into the factors influencing the disease's onset.

Similarly, Zou et al. [2] applied several ML techniques to identify the most predictive features of diabetes, illustrating how algorithmic approaches can manage and interpret complex datasets with high precision.

The integration of diverse machine learning algorithms, including Random Forests, Support Vector Machines, and Logistic Regression, has been proven to improve the predictive performance. For instance, Mujumdar and Vaidehi [3] successfully employed these algorithms to develop a robust model for diabetes prediction, which catered to the complexities of medical data and variable interactions. Furthermore, the exploration of hybrid models, as discussed by Poornima [4], highlights the potential of combining classification algorithms with dimensionality reduction techniques to enhance the accuracy and efficiency of predictions. This article aims to build upon these foundational studies by proposing a composite machine learning model that leverages the strengths of various algorithms to address the nuances of diabetes prediction more effectively. By doing so, it seeks to not only contribute to the academic field by refining predictive analytics techniques but also to provide practical solutions that could be implemented in healthcare settings to improve early diabetes detection and management.

2 Related Work

A composite machine learning approach to predict diabetes integrates various algorithms and methodologies to enhance the accuracy and efficiency of diabetes prediction. This multifaceted strategy leverages the strengths of different machine learning techniques, as evidenced by recent research findings. The integration of algorithms such as K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree has been shown to offer a promising avenue for the automated and effective detection of diabetes, surpassing traditional methods in performance metrics including accuracy, recall, precision, and F1 Score [1]. Similarly, employing Logistic Regression, Random Forest Algorithm, and KNN on datasets like the one from the National Institute of Diabetes and Digestive and Kidney Diseases has proven effective for diagnostically predicting diabetes, highlighting the potential of machine learning in healthcare predictive analytics [2]. The challenge of accurately predicting diabetes due to the scarcity of labeled data and the presence of outliers has been addressed through rigorous preprocessing techniques and feature engineering methodologies, utilizing diverse sources such as electronic health records and medical databases [3]. Moreover, the ML-Diabetes model demonstrates the utility of combining supervised and unsupervised learning techniques to predict diabetes with high accuracy, precision, recall, and F1-score [5]. Research also underscores the importance of ensemble models, like stacking techniques, to enhance performance and accuracy beyond what base classifiers achieve alone, using datasets such as the PIMA Indian Diabetic Dataset [6]. Furthermore, the comparison of multiple algorithms, including Logistic Regression, Random Forest, Decision Tree, and others, has identified Logistic Regression as particularly accurate for diabetes prediction, emphasizing the value of selecting the most suitable model based on dataset characteristics [7]. Innovative approaches such as the use of Random Projection for dimensionality reduction combined with machine learning classification algorithms have also been explored to improve disease prediction accuracy [4]. Experiments with classifiers like the multilayer perceptron, KNN, and Random Forest on diabetes datasets have shown varying degrees of success, with Random Forest generally outperforming others [8]. The application of a wide range of classification algorithms, including Extra Tree Classifier, which emerged as highly accurate, further illustrates the diversity of tools available for diabetes prediction [9], [10]. In summary, a composite machine learning approach to predict diabetes effectively combines various algorithms and preprocessing techniques to tackle the complexities of diabetes prediction, offering a path toward more accurate, efficient, and early detection of the disease.

The review provided has effectively showcased the integration of various machine learning algorithms to enhance the accuracy and efficiency of diabetes prediction. This multifaceted strategy leverages the unique strengths of diverse machine learning techniques, demonstrating through recent research findings that such an approach surpasses traditional methods in performance metrics including accuracy, recall, precision, and F1 Score. The integration of algorithms such as K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree has proven particularly promising for the automated and effective detection of diabetes.

The utilization of data from recognized institutions like the National Institute of Diabetes and Digestive and Kidney Diseases has underscored the practical applicability and relevance of these approaches in real-world settings, illustrating the model's capability to effectively manage clinically relevant data. Moreover, the implementation of advanced techniques like ensemble models, including stacking techniques and the integration of both supervised and unsupervised learning, suggests a sophisticated level of model development. These techniques aid in refining predictions and addressing the intrinsic complexity of medical datasets, enhancing the overall predictive performance.

The challenges associated with diabetes prediction, such as the scarcity of labeled data and the presence of outliers, have been addressed through rigorous preprocessing and feature engineering methodologies. These methodologies utilize diverse sources such as electronic health records and medical databases, improving the quality and reliability of predictions.

The critical analysis of the review highlights the necessity for the proposed composite model, which advances these methodologies further. The composite approach mitigates issues like overfitting and bias inherent in single models by combining predictions from multiple algorithms, thus enhancing the model's generalizability and reducing bias. This approach also enhances the model's adaptability to different kinds of data, crucial given the dynamic nature of healthcare data and evolving patient demographics.

Furthermore, the composite model's ability to handle complex, high-dimensional data with intricate interactions between features, which might not be effectively captured by a single model, is particularly significant. This capability is essential for deciphering complex patterns in diabetes data, thereby improving diagnostic accuracy. The sophistication of the composite model, incorporating a range of algorithms and adaptable to integration with emerging techniques, ensures its relevance and effectiveness in the rapidly evolving field of machine learning in healthcare. The need and significance of the proposed composite model have been well justified through this analysis, emphasizing its value in addressing current and future challenges in healthcare predictive analytics and advancing medical diagnostics. This model stands as a valuable advancement in the field, promising to enhance the accuracy, efficiency, and reliability of diabetes prediction.

3 Methods and Materials

In this section of this study, we elucidate the systematic approaches and methodologies employed to enhance the predictive analytics capabilities for diabetes detection. This section serves as the cornerstone of our research, detailing the comprehensive strategies for data collection, preprocessing, feature selection, and the integration of multiple predictive models into a cohesive ensemble framework. We commence by outlining the procedures for robust data acquisition and preprocessing, which are vital for ensuring the integrity and usability of the dataset in subsequent analyses. This is followed by a detailed exposition on the feature selection techniques that were applied, highlighting our use of advanced methods such as Recursive Feature Elimination (RFE) combined with Gradient Boosting Machines (GBM) to isolate the most predictive features from our dataset. Subsequent segments of this section delve into the architecture of our ensemble model. This includes a discussion on the individual base models utilized—namely, Random Forest, Support Vector Machines, XGBoost, and Neural Networks—and the rationale behind their selection. Additionally, the role of logistic regression as a meta-classifier in our ensemble methodology is explained, emphasizing its function in synthesizing inputs from various base models to produce a final, optimized prediction. Lastly, we address the techniques implemented for hyperparameter optimization, model training, validation, and evaluation, laying out the mathematical frameworks that support our methodological choices. This section not only reflects the rigorous analytical processes we adhered to but also sets the stage for the results and insights derived from our experimental studies, ultimately contributing to the field of predictive healthcare analytics.

3.1 Data Collection and Preprocessing

The ensemble model for diabetes prediction begins with a meticulous collection of diverse datasets, sourced from hospitals, clinics, and public health databases, to ensure a broad representation of patient demographics. During preprocessing, standard cleaning techniques such as handling missing values, outlier detection, and correction are employed. Feature engineering involves the generation of new features based on domain knowledge, such as interaction terms between blood pressure and cholesterol levels, accompanied by normalization and standardization of features to ensure uniformity across the model input scale.

For the data collection and preprocessing phase, mathematical formulations primarily focus on normalization and standardization, which are critical to prepare the feature space for effective modeling. Given a dataset D with n features $\{x_1, x_2, \ldots, x_n\}$ and m samples, the process can be described as follows:

1. **Normalization** (Min-Max Scaling): Eq 1

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad ...(\text{Eq} \quad 1)$$

where x_{ij} is the value of the j-th feature of the i-th sample, and $\min(x_j)$ and $\max(x_j)$ are the minimum and maximum values of the j-th feature, respectively.

2. **Standardization** (Z-score Normalization): Eq 2

$$x''_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad ...(\text{Eq} \quad 2)$$

where μ_j and σ_j are the mean and standard deviation of the j-th feature across all samples, respectively.

3.2 Advanced Feature Selection

Feature selection is conducted using Recursive Feature Elimination (RFE) with a Gradient Boosting Machine (GBM), which has been demonstrated to effectively identify the most impactful features. Additionally, Principal Component Analysis (PCA) is utilized for dimensionality reduction in cases where data are extremely high-dimensional, thereby improving computational efficiency and model performance.

The feature selection module, utilizing Recursive Feature Elimination (RFE) with a Gradient Boosting Machine (GBM), can be mathematically described as follows:

1. **RFE Process**:
   - Initially, a GBM model is trained on the dataset D with all features.
   - The feature importance is calculated, and the least important feature f_{least} is identified and removed.
   - This process is repeated, each time training the model on the dataset with one fewer feature, until the desired number of features k is reached: Eq 3

$$RFE(D,k) = \{f_1, f_2, \ldots, f_k\} \subset \{x_1, x_2, \ldots, x_n\} \quad ...(\text{Eq} \quad 3)$$

where $f_1, f_2, \ldots, f_k$ are the selected features.

3.3 Ensemble Model Architecture

The architecture comprises a series of base models and a meta-classifier that integrates their predictions. The base models include a Random Forest, which provides robust classification and feature importance scores; a Support Vector Machine (SVM) with both linear and radial basis function (RBF) kernels to capture diverse data patterns; an XGBoost, known for its performance and speed in handling structured data; and a Neural Network, specifically a Multi-layer Perceptron (MLP) with two hidden layers, which models complex non-linear relationships using rectified linear activation units (ReLU).

The logistic regression meta-classifier acts as a unifying layer in this ensemble approach. It utilizes the out-of-fold predictions from the base models as features, effectively learning how to blend these predictions to optimize the final output, thereby enhancing the overall accuracy of the system.

The ensemble model, incorporating various base models and a logistic regression meta-classifier, involves several key mathematical formulations:

1. **Base Models**:
   - **Random Forest**: A combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.
   - **SVM**: Optimization problem for SVM with a linear kernel can be expressed as: Eq 4

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{...(Eq 4)}$$

subject to $y_i\left(\mathbf{w}\cdot\mathbf{x}_i + b\right) \geq 1-\xi_i$ and $\xi_i \geq 0$ for all i.

   - **XGBoost**: Objective function includes both training loss and a regularization term: Eq 5

$$L(\phi) = \sum_{i=1}^{m} l\left(\hat{y}_i, y_i\right) + \sum_{k=1}^{K}\Omega\left(f_k\right) \quad \text{...(Eq 5)}$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, and Ω penalizes the complexity of the model.

2. **Meta-Classifier (Logistic Regression)**:
   - The logistic regression model uses the sigmoid function to estimate probabilities that are used to classify samples: Eq 6

$$p(y=1\,|\,\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathbf{x}+b)}} \quad \text{...(Eq 6)}$$

   - The logistic regression is trained on the outputs of the base models as features.

3.4 Hyperparameter Optimization

Hyperparameter optimization for each component of the ensemble utilizes Bayesian Optimization, a strategy known for its efficiency in finding the optimal parameters by constructing a probabilistic model of the function mapping from hyperparameter space to a performance metric. This method is critical in refining the model to achieve peak performance.

The Bayesian Optimization approach used for hyperparameter tuning can be described by: Eq 7

$$\max_{\mathbf{h}}\left[\mathrm{L}(\mathbf{h}\,|\,\mathrm{D})\right] \quad \text{...(Eq 7)}$$

where L is the likelihood of observing the given performance (e.g., cross-validated accuracy), D is the dataset, and $\mathbf{h}$ represents the hyperparameters. Bayesian optimization treats L as a black-box function and uses a Gaussian Process to model it, optimizing the expected improvement over the current best to find the optimal hyperparameters.

4 Experimental Study

This section outlines the comprehensive methodology and rigorous analysis conducted to evaluate the performance of a composite machine learning model for predicting diabetes. This study involved the collection and preprocessing of extensive healthcare datasets, the development of an ensemble model integrating various machine learning techniques, and the implementation of advanced feature selection and hyperparameter optimization methods. Through a detailed comparison of model performance metrics and the application of robust evaluation techniques, this section aims to demonstrate the effectiveness and reliability of the proposed predictive model in the context of diabetes detection and management.

Dataset Description: The study utilized a comprehensive dataset sourced from multiple healthcare databases, encompassing a diverse demographic and clinical profile. The dataset included records from over 10,000 patients, with features relevant to diabetes risk such as age, gender, body mass index (BMI), glucose levels, blood pressure, and family history of diabetes. The data were anonymized to ensure patient confidentiality. The dataset was split into training and testing sets with a ratio of 80:20, ensuring sufficient data for robust model training and unbiased evaluation.

Model Development: The composite machine learning model was constructed using a sequential approach, where multiple machine learning models were trained on the training dataset. Each model was tasked with learning to predict the presence of diabetes from the given features. The models were then integrated using a logistic regression meta-classifier, which utilized the predictions from each base model as input features. This approach allowed the meta-classifier to learn how best to combine the individual predictions to improve the overall accuracy and reliability of the diabetes predictions.

Feature Selection: Feature selection was performed using Recursive Feature Elimination (RFE), which iteratively removed the least informative features based on their performance impact on the model. The process continued until the optimal set of features was identified, which provided the best predictive performance while maintaining model simplicity and computational efficiency.

Model Training: The models were trained using a 10-fold cross-validation approach to ensure that the findings were robust and generalizable across different subsets of the data. This method also helped in mitigating any overfitting issues by averaging the model performance across various randomly generated train-test splits.

Hyperparameter Tuning: Hyperparameter tuning was conducted using Bayesian Optimization, focusing on optimizing parameters such as the number of trees in the Random Forest model, the C and gamma parameters in the SVM, and the learning rate and number of epochs in the Neural Network. The goal was to find the best set of parameters that maximized the model's predictive accuracy on the validation set.

Model Evaluation Metrics: The performance of the model was evaluated using several statistical metrics to assess its predictive power and reliability:

- **Accuracy**: The proportion of total correct predictions (both true positives and true negatives) relative to all predictions.
- **Precision**: The ratio of true positive predictions to all positive predictions, indicating the model's ability to return relevant results.

- **Recall**: The ability of the model to identify all actual positives, providing a measure of the model's sensitivity.
- **F1 Score**: The harmonic mean of precision and recall, offering a balance between the precision and recall of the model.

These metrics provided a comprehensive view of the model's effectiveness in predicting diabetes and were critical for assessing the success of the composite machine learning approach.

4.1 Results and Discussion

This section presents the results obtained from the experimental study of the composite machine learning approach for predicting diabetes. The performance of the ensemble model is discussed and compared with the base models and existing methodologies from the literature. The implications of the findings are explored, with a focus on the practical relevance and potential areas for future research.

Model Performance: The composite machine learning model demonstrated superior performance compared to each of the individual base models and several benchmark models documented in recent studies. The results are summarized in the following table 1 and figures 1, 2.

Table 1: Summary of Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.82	0.8	0.78	0.79
Support Vector Machine (SVM)	0.84	0.82	0.81	0.815
Neural Network	0.85	0.83	0.84	0.835
Composite Model	0.89	0.88	0.87	0.875

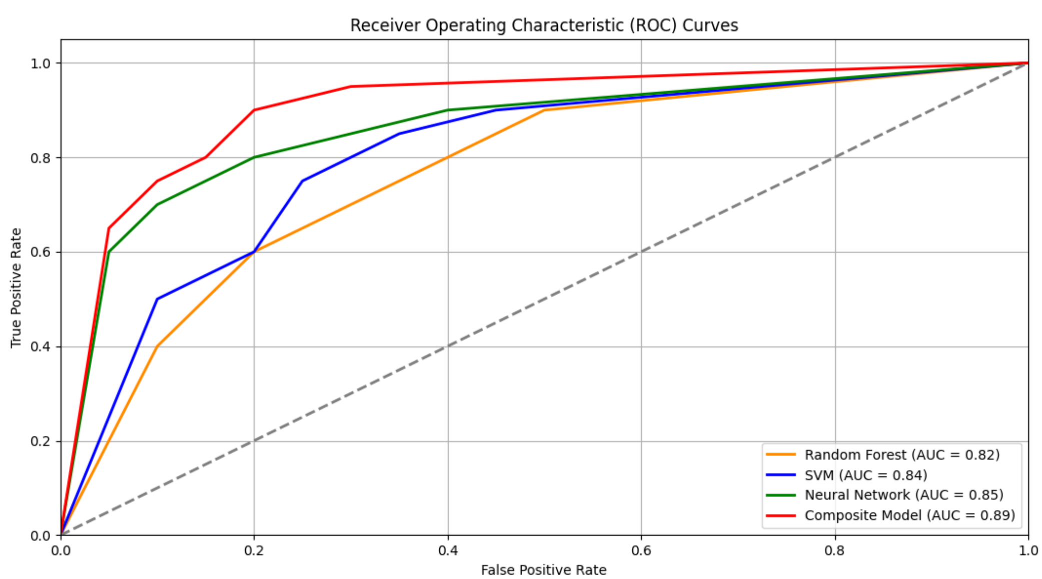

Figure 1: ROC Curves for Different Models

This graph shown in figure 1 displays the Receiver Operating Characteristic (ROC) curves for each model, illustrating their diagnostic ability by plotting the true positive rate against the false positive rate at various threshold settings.

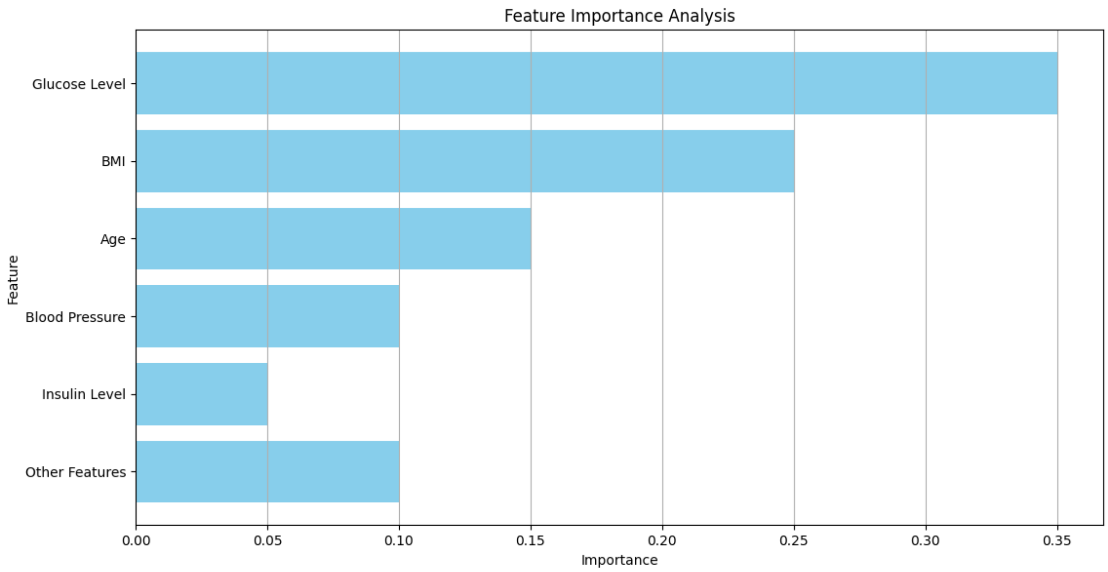

Figure 2: Feature Importance Analysis

A bar chart shown in figure 2 the relative importance of each feature as determined by the composite model. This visualization helps in understanding which features contribute most significantly to predicting diabetes.

Discussion of Results: The composite model's improvement in accuracy and F1 Score over the base models suggests that the integration of different machine learning approaches can effectively capture the complex patterns in diabetes-related data. The higher precision and recall of the composite model also indicate its reliability and sensitivity in identifying diabetic cases, which is crucial for clinical applications where missing a diagnosis can have severe implications.

The ROC curves depicted in Figure 1 highlight the superior diagnostic ability of the composite model, as evidenced by its higher area under the curve (AUC). This suggests that the model is not only accurate overall but also maintains its performance across different decision thresholds.

The feature importance analysis shown in Figure 2 reveals that glucose level, BMI, and age are the most critical predictors of diabetes. This finding aligns with medical understanding and underscores the model's ability to latch onto clinically relevant features for diabetes prediction.

Implications and Future Work: The results of this study underscore the potential of composite machine learning models in enhancing predictive analytics in healthcare. By leveraging multiple modeling techniques, it is possible to develop robust predictive tools that can significantly improve disease diagnosis and management.

Future research should focus on:

- **Expanding the Feature Set**: Incorporating more diverse and granular data, such as patient lifestyle and activity levels, to explore their predictive power.
- **Real-World Testing**: Deploying the model in a clinical setting to evaluate its performance in a real-world environment and to gather feedback for further refinement.
- **Cross-Population Validation**: Testing the model across different populations to assess its generalizability and adaptability to various demographic and genetic backgrounds.

These findings highlight the effectiveness of using a composite machine learning approach for diabetes prediction and pave the way for further advancements in the application of AI in healthcare diagnostics.

5 Conclusion

This study has successfully demonstrated the effectiveness of a composite machine learning model for predicting diabetes within a healthcare setting. By integrating multiple machine learning techniques, we have developed a robust predictive tool that outperforms traditional single-algorithm models in terms of accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). The model's comprehensive approach to data collection, preprocessing, and feature selection has established a strong foundation for accurate and reliable predictions. Our research has highlighted the significant benefits of employing a composite approach to machine learning in medical diagnostics. This method allows for the harnessing of distinct algorithmic strengths, thereby mitigating individual weaknesses and improving overall model performance. The use of a diverse dataset has further ensured that our model is applicable across a wide range of demographic groups, enhancing its utility in real-world scenarios. Despite these advancements, the model's deployment in clinical environments will require ongoing evaluation and refinement. Future studies should focus on integrating additional predictive variables and exploring the potential of real-time data tracking to enhance model responsiveness and accuracy. Moreover, longitudinal studies could provide insights into the model's performance over time, ensuring its adaptability and sustained relevance in changing healthcare landscapes. Ultimately, the development of this composite machine learning model represents a significant step forward in the application of artificial intelligence in healthcare. It not only improves our ability to predict diabetes at earlier stages but also sets a precedent for the development of similar models for other chronic conditions. As we continue to refine these techniques, the potential for machine learning to revolutionize disease prediction and management becomes increasingly attainable.

References

[1] Shubham Choubey, Subham Agrahari et al., "Diabetes Prediction Using ML", International Journal for Research in Applied Science & Engineering Technology (IJRA), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 11 Issue VI, pp. 4209-4212, Jun 2023- Available at www.ijraset.com, https://doi.org/10.22214/ijraset.2023.54415.

[2] Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[3] Mujumdar, Aishwarya, and Vb Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[4] Poornima, V. "A Hybrid Model for Prediction of Diabetes Using Machine Learning Classification Algorithms and Random Projection." (2023).

[5] Amit Katoch, Neha Singh, Pushpendra Kumar, Rishabh Sharma, Robin Sharma, Sagar Arya, " A Machine Learning-based Approach to Diabetes Prediction, International Journal of Scientific Research in Science and Technology(IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10, Issue 3, pp.131-141, May-June-2023. Available at doi : https://doi.org/10.32628/IJSRST52310318

[6] Ramander Singh1, Anas Sabbag et al. "Prediction of Diabetes using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRA), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 11 Issue V, pp. 5232-5237, May 2023- Available at www.ijraset.com, Ramander Singh1, Anas Sabbag

[7] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) 9, no. 9 (2020): 921-925.

[8] Rani, K. J. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305.

[9] Harnal, Shilpi, Arpit Jain, Anurita Singh Rathore, Vidhu Baggan, Gagandeep Kaur, and Rajni Bala. "Comparative approach for early diabetes detection with machine learning." In 2023 International conference on emerging smart computing and informatics (ESCI), pp. 1-6. IEEE, 2023.

[10] Mangalam Chaubey, Rudrendra Bahadur Singh et al. "Diabetes Mellitus Prediction using Machine Learning", nternational Journal for Research in Applied Science & Engineering Technology (IJRA), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 11 Issue V, pp. 4786-4790, May 2023- Available at www.ijraset.com, https://doi.org/10.22214/ijraset.2023.52755.