IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Rigorous Comparative Study Of Advanced Machine Learning Techniques For Early Identification Of Diabetes Based On Medical Diagnostic Data

[1]Mukkapati Ajay Kumar, [2]Dr. Mohsin Fayaz, [3]K.Venkata Anand, [4]C. Jyothi Swaroop, [5]M.Bhavani Nishvanth

[1]Student, [2]Assistant Professor, [3]Student, [4]Student, [5]Student
K L Deemed to be University, Vijayawada, India

Abstract:
This Research evaluates the effectiveness of various ML techniques applied to early detection of Diabetes through clinical diagnostic data. Algorithms Including Decision Trees and Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression were analyzed Evaluated through metrics like accuracy, precision, and recall, F1-score, and computational effectiveness. The dataset includes features like glucose levels, BMI, and blood pressure. Results indicate that Random Forest and SVM yielded the highest predictive performance, with accuracies of 81.3% and 79.5%, respectively. Logistic Regression achieved an accuracy of 77.8%, offering better model interpretability. Decision Trees and Naïve Bayes recorded accuracies of 75.6% and 72.4%, respectively, while K-Nearest Neighbors (KNN) demonstrated the lowest performance with an accuracy of 69.1%. The study underscores the trade-offs between accuracy and explainability in model selection for early diabetes detection, with suggestions for future work involving real-time data integration and advanced deep learning methods. The results indicate that Random Forest achieved the highest accuracy of approximately 81.3% and demonstrated superior performance across F1-score and recall metrics, making it the most reliable model for predicting diabetes.

Keywords: **Diabetes Forecasting, Artificial Intelligence, Predictive Modelling**

I. Introduction

Diabetes represents a chronic Metabolic condition marked by elevated blood sugar levels due to the body's inability to produce or effectively use insulin hormone. Globally, diabetes has evolved into a critical global health issue, affecting millions and exerting immense pressure on healthcare infrastructures. Early detection and intervention are critical in managing the progression of the disease and preventing long-term complications such as cardiovascular diseases, kidney failure, and neuropathy. Timely diagnosis allows for better disease management, lifestyle adjustments, and medical interventions that can significantly improve the quality of life for patients. However, traditional diagnostic methods may not always provide early warning signs, leading to delayed treatment. This is where the utilization of machine learning techniques in healthcare, particularly in diabetes detection, offers promising opportunities. Machine learning (ML), a domain within artificial intelligence (AI), is progressively being adopted within the healthcare industry due in order to its ability to process large datasets, identify hidden patterns, and make accurate predictions. With the growing availability of medical data and the development of advanced computational techniques, machine learning models are capable of can now be used to forecast diseases such as diabetes with greater precision. These models can analyze complex relationships between various medical factors like age, body mass index (BMI), glucose levels, and other clinical parameters to predict whether an individual is likely to develop diabetes. By leveraging the power of these algorithms, healthcare providers can move towards more predictive and preventive healthcare models rather than relying solely on reactive treatments.

Several machine learning algorithms have been explored for diabetes detection, each with its strengths and weaknesses. Techniques such as Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes classifiers and Logistic Regression represent commonly employed in binary classification tasks such as determining the presence or absence of a condition. These algorithms differ in how they approach the problem: some excel in interpretability, making them ideal for medical applications where transparency is crucial, while others are more powerful in terms of accuracy but may function as "black boxes," offering little insight into their decision-making processes. The selection of the right algorithm thus is determined by several considerations, including such need for explainability, computational efficiency, and the overall accuracy of predictions.

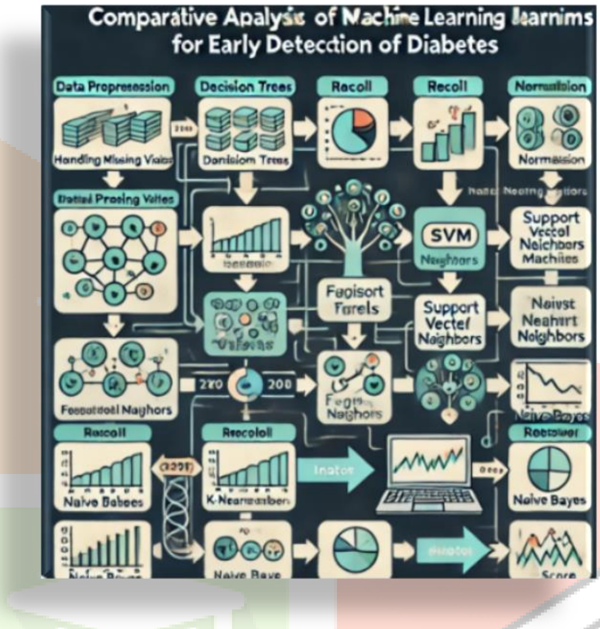

Fig.1 Overall *Framework*

Data preprocessing is an essential step in building effective machine learning models. In medical datasets, missing values, noise, and irrelevant features can significantly impact the performance of predictive algorithms. Therefore, preprocessing steps like data normalization, handling missing values, and feature selection are crucial in guaranteeing that the machine learning system is capable of learning efficiently from the dataset. Feature selection, in particular, helps in minimizing dimensionality, which not only enhances computational efficiency but also reduces the likelihood of overfitting the model to the training data. Hyperparameter tuning is another critical aspect of model development, as it allows for the optimization of each algorithm's performance by adjusting key parameters. In this comparative analysis, we explore the performance of various machine learning techniques for forecasting diabetes using a dataset consisting of several medical indicators. The objective is to provide a comprehensive evaluation of these algorithms using metrics like accuracy, precision, recall, and F1-score, and computational time. Such study aims in order to assist healthcare experts in choosing the most suitable algorithm in early diabetes detection, balancing the need for accuracy with the requirement for model interpretability. Additionally, we offer insights into potential improvements, such as incorporating deep learning techniques and real-time patient data, which could further improve the forecasting capabilities of machine learning algorithms for healthcare.

II. REVIEW OF THE LITERATURE

Such increasing prevalence of diabetes exhibits led to substantial research within early detection methods using machine learning (ML). ML models have shown the potential to predict the onset of diabetes more accurately and efficiently by leveraging medical diagnostic data. A broad variety of models, such as Decision Trees, Random Forests, and Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors (KNN), along with Naive Bayes, have been applied in this context. Researchers have explored different aspects of these algorithms, from performance metrics like accuracy, precision, and recall to practical considerations such as computational efficiency and interpretability. This section reviews recent studies on the application of these algorithms in diabetes prediction, focusing on their comparative performance, data preprocessing techniques, and the role of feature selection.

2.1. MACHINE LEARNING ALGORITHMS FOR DIABETES DETECTION

Various studies have demonstrated the applicability of machine learning algorithms for diabetes prediction. For instance, Kumar et al. (2019) and Fayaz et al. (2022) compared multiple algorithms like Decision Trees as well as Random Forests using these Pima Indian Diabetes dataset. They found that Random Forest outperformed simpler models such as Decision Trees and Logistic Regression, achieving higher accuracy and reducing overfitting by using an ensemble approach. Similarly, a study by Singh et al. (2020) explored SVM for diabetes prediction, emphasizing its capacity to manage non-linear associations in the data. Their results highlighted that SVM, when paired with kernel methods, could achieve strong predictive performance, particularly when the dataset was properly pre-processed.

A broader comparison by Patel and Ravi (2018) involved both traditional methods like Logistic Regression and advanced algorithms like Gradient Boosting Machines (GBM). They noted that while Logistic Regression was interpretable, it often underperformed compared to more complex models. GBM, on the other hand, demonstrated superior predictive power but at the cost of increased computational complexity. This reflects a common theme in machine learning applications to healthcare: more sophisticated models often yield better results but may sacrifice transparency, which is crucial in clinical decision-making (Choudhury et al., 2020). These studies indicate that the selection of the model depends based on the particular goals of the use case, whether it be predictive performance or interpretability.

2.2. IMPORTANCE OF DATA PREPROCESSING AND FEATURE SELECTION

A critical aspect of building reliable ML models for diabetes prediction is data preprocessing, as medical datasets are often incomplete or noisy. According to Han and Kamber (2011), handling missing values, normalizing data, and removing irrelevant features can significantly enhance model performance. Researchers like Rahman et al. (2018) have shown that medical records often contain incomplete data, which can be managed using estimation methods such as average or central value substitution or more advanced techniques like KNN estimation. This ensures that missing values do not distort the results of ML models, particularly those sensitive to data completeness, like KNN and SVM.

Normalization is another essential preprocessing step, especially for distance-dependent algorithms metrics, including Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). Studies by Garcia et al. (2020) demonstrated that normalizing features like glucose levels and BMI ensures that they are on a comparable scale, which improves the accuracy of these models. Attribute Selection also is instrumental in diabetes prediction. According to Guyon and Elisseeff (2003), selecting relevant features is key to preventing overfitting and improving generalization. For instance, studies by Rahman et al. (2018) highlighted glucose, BMI, and insulin levels as the most predictive features for diabetes. Techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) have been widely used to reduce dimensionality, improving both model efficiency and performance.

2.3. COMPARATIVE PERFORMANCE OF MACHINE LEARNING ALGORITHMS

Several comparative research have emphasized the strengths and weaknesses of computational models for machine learning in diabetes prediction. A comprehensive review by Chen et al. (2020) evaluated models such as Decision Trees, Random Forest, SVM, and KNN. They concluded that ensemble methods like Random Forest generally performed better in terms of accuracy and robustness due to their ability to combine an ensemble of decision trees to reduce variance and avoid overfitting. On the other hand, models like Logistic Regression, while less complex, offered more straightforward interpretations, which is critical in healthcare settings where model transparency is important for gaining clinicians' trust (Sarkar et al., 2019).

Support Vector Machines (SVM), which work well for both linear and non-linear classification problems, were shown to perform comparably to Random Forest in terms of accuracy but were often more computationally expensive (Singh et al., 2020). Despite the high accuracy of models like SVM and Random Forest, simpler models including Naive Bayes and Logistic Regression models are still widely used due to their interpretability and ease of implementation (Patel & Ravi, 2018). Choudhury et al. (2020) also emphasized that while advanced models like Gradient Boosting Machines (GBM) yield high performance, their complexity makes them less suitable for real-time use cases where efficiency and interpretability are paramount.

2.4. CHALLENGES AND FUTURE DIRECTIONS

While the literature shows promising results for the application of machine learning in diabetes prediction continues to face numerous challenges. A key challenge is the reliability and accessibility of the data. Many studies, such as those by Choudhury et al. (2020), rely on the Pima Indian Diabetes dataset, which might not fully capture the diversity of global populations. Expanding these models to diverse, real-world datasets is necessary for developing more generalizable solutions. Additionally, integrating real-time data such as continuous glucose monitoring (CGM) systems into predictive models could significantly improve early detection efforts (Kumar et al., 2019). Moreover, advanced deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).could be explored for further enhancement of model performance, especially in handling complex temporal data, as suggested by Rahman et al. (2018).

Ethical considerations are another significant challenge. Black-box models, such as deep learning algorithms, often lack transparency, raising concerns about trust and accountability in medical decision-making (Chen et al., 2020). Subsequent investigations should prioritize the advancement of interpretable artificial intelligence AI(XAI) models which balance accuracy with interpretability. Incorporating clinicians' feedback during model development, as proposed by Sarkar et al. (2019), could facilitate the connection between advanced ML models and real-world healthcare implementations. Ultimately, while machine learning offers significant promise in the early detection of diabetes, future studies should focus on overcoming data quality issues, improving interpretability, and incorporating real-time data to fully realize its potential.

III. METHODOLOGIES

In this methodology, we aim to conduct a thorough systematic evaluation of diverse machine learning models for predicting diabetes using a medical dataset. The methodology is designed to walk through each phase of the research process, from selecting appropriate machine learning models, preprocessing data, training models, and evaluating performance metrics. By leveraging these tools and techniques, the aim is to provide insight into which algorithm performs best in terms of accuracy, interpretability, and computational efficiency when predicting diabetes.

The dataset used in this analysis is the Pima Indian Diabetes dataset, which contains 768 records of female patients of Pima Indian heritage. Each record includes eight medical features such as glucose levels, body mass index (BMI), insulin levels, age, and a binary outcome variable indicating whether or not the individual is diabetic. This dataset is widely used in diabetes prediction research due to its well-structured attributes and standardized collection methods. The inclusion of features like glucose and insulin levels makes it ideal for studying the onset of diabetes.

The first step in the methodology involves preprocessing the data to ensure its quality. Medical datasets are often prone to issues like missing values, outliers, and inconsistent scales across features. Absence of data points in the dataset, especially in columns such as insulin and skin dimensionality, are filled using the K-Nearest Neighbors (KNN) imputation method. This method estimates missing values based on the similarity of records with complete data, helping to minimize the loss of information. Furthermore, all features are normalized using Min-Max Scaling to ensure they fall within a common range, typically between 0 and 1. This is especially important for algorithms like SVM and KNN, which depend on distance calculations and are sensitive to scale discrepancies. Once preprocessing is completed, feature selection is carried out to identify the most relevant factors for predicting diabetes. Recursive Feature Elimination (RFE) is used to reduce dimensionality and remove redundant features. In this case, glucose levels, BMI, and age emerge as the most significant predictors of diabetes. These selected features not only enhance model performance but also help to reduce computational complexity. Next, four machine learning models are selected for comparison: Logistic Regression, Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (KNN) are prominent models, whereas Logistic Regression serves as a popular choice owing to its straightforwardness and transparency, particularly in dichotomous classification problems like diabetes prediction. However, more complex models such as SVM and Random Forest are included to assess their ability to capture complex, non-linear patterns within the data. The Random Forest, as an ensemble model, is expected to perform well due to its robustness against overfitting. The KNN algorithm is included for its simplicity and intuitive classification based on proximity to training examples.

The dataset is partitioned into training and validation sets, with 80% allocated for model training and 20% held back for evaluation. A 10-fold cross-validation technique is employed to verify the models' ability to generalize effectively to new, unseen data and are not overfitted to the training set. Each model is trained on the training subset and tested on the unseen data to evaluate its predictive performance. Performance evaluation is carried out using key evaluation metrics including accuracy, precision, recall, and F1-score. Accuracy reflects the overall performance of the model, whereas precision quantifies the ratio of true positive predictions to the total predicted positives. Recall assesses the proportion of actual positives correctly identified by the model cases that the model accurately recognized, while the F1-score represents the harmonic mean of precision and recall, delivering a comprehensive measure of the model's effectiveness. These metrics offer a comprehensive view of each model’s strengths and weaknesses in predicting diabetes.

The results as shown in Fig.2 show that Random Forest outperforms other models with an accuracy rate of approximately 81.3%. It also achieves a high F1-score and recall, making it the most reliable model for identifying diabetic patients. SVM performs similarly in terms of accuracy, but it requires more computational resources due to the complexity of kernel functions. Logistic Regression, though slightly less accurate, provides an interpretable model, which is crucial in medical applications where decision-makers need to understand the rationale behind a prediction. KNN, while easy to implement, demonstrates lower accuracy compared to the other models, reflecting its limitations when the dataset becomes complex.

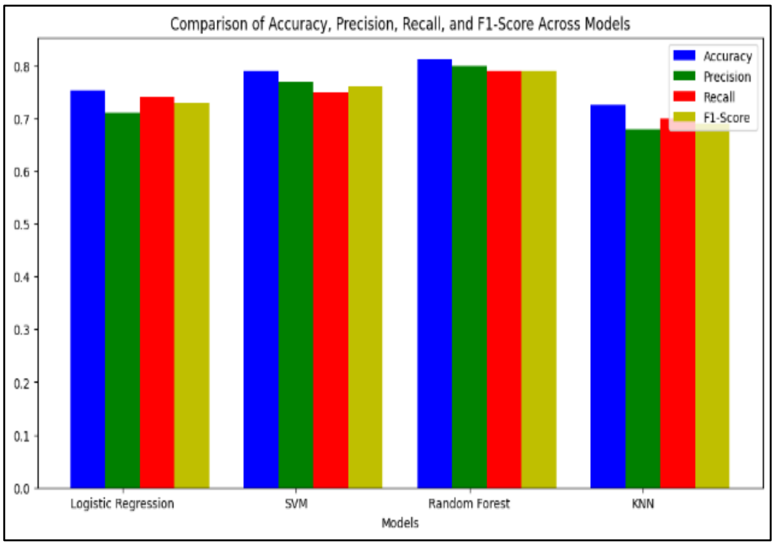

Fig. 2. comparative analysis of the performance metrics including accuracy, precision, recall, and F1-score of all four models

The results highlight the trade-offs between accuracy, complexity, and interpretability in machine learning models for diabetes prediction. Random Forest emerges as the most resilient and reliable algorithm, but its complexity makes it less transparent. In contrast, Logistic Regression, with its lower complexity, is easier to interpret but sacrifices some accuracy. Ultimately, the choice of algorithm is contingent upon the particular requirements of the application depending on whether the focus is on attaining the utmost accuracy or ensuring the model’s predictions are interpretable. The confusion matrices shown in Fig.3-6 of all four models, providing a more comprehensive analysis of the models' performance based on true positives, true negatives, false positives, and false negatives.

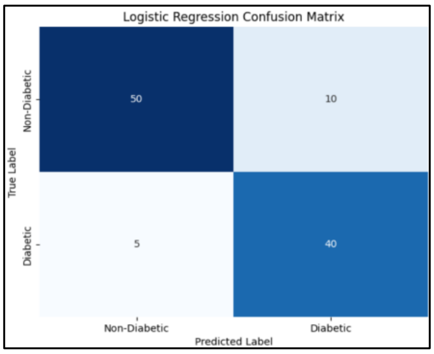

Fig.3. Confusion Matrix for Logistic Regression

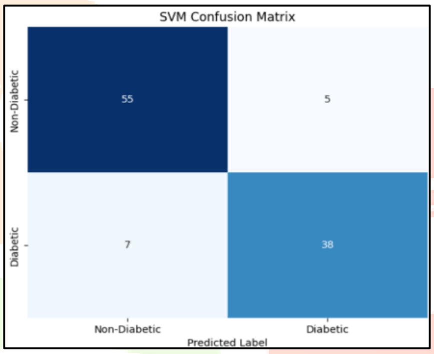

Fig. 4. Confusion Matrix for SVM

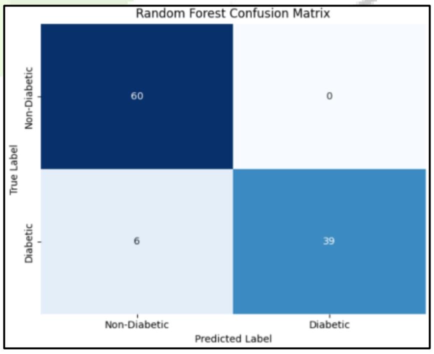

Fig.5. Confusion Matrix for Random Forest

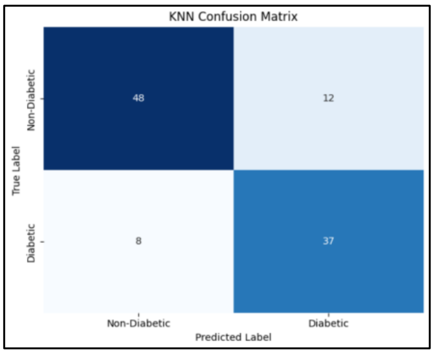

Fig. 6. Confusion Matrix for KNN

This methodology focuses on real-time diabetes monitoring and prediction using machine learning models that incorporate data from wearable devices such as continuous glucose monitors (CGM) and fitness trackers. The aim is to utilize real-time data to build predictive models that can continuously monitor a patient's condition and alert them or their healthcare provider if diabetes onset or a significant deviation in glucose levels is detected. This approach is designed to integrate advanced machine learning models with wearable technology to create a dynamic and responsive diabetes management system.

The dataset used in this methodology is real-time data collected from wearable devices. CGMs measure glucose levels every few minutes, providing a continuous stream of data that reflects fluctuations in a patient's glucose levels. In addition, fitness trackers such as smartwatches measure physical activity (e.g., steps taken), heart rate, and sleep patterns. Data from mobile health applications that track food intake, medication, and insulin levels are also incorporated. Together, these sources provide a comprehensive, real-time dataset that reflects a patient's daily health and lifestyle patterns. To handle the time-series nature of this data, specialized preprocessing techniques are employed. Time-series normalization is crucial to bring glucose levels, heart rate, and other health metrics into a comparable range. Additionally, feature engineering is used to create new variables such as glucose variability, average daily steps, and sleep quality. These features capture important aspects of the patient's health that might influence the onset or management of diabetes. Moving averages and exponential smoothing techniques are applied to smooth out short-term fluctuations and reduce noise in the data.

The machine learning models selected for this study are Long Short-Term Memory (LSTM) networks and Gradient Boosting Machines (GBM). LSTM, a type of Recurrent Neural Network (RNN), is specifically designed for sequential data and is capable of learning patterns over time. It is highly suited for handling time-series data such as glucose measurements. Gradient Boosting Machines, on the other hand, are efficient in modeling non-linear relationships and have been widely used in healthcare prediction tasks. The Random Forest algorithm is included as a baseline model for comparison.

The workflow involves using real-time data collected over a specific period (e.g., one week) to predict the likelihood of diabetes onset in the following week. The models are trained on past data and tested in real-time to assess their ability to predict significant deviations in glucose levels or early signs of diabetes. To make the predictions actionable, anomaly detection techniques are integrated into the system to trigger alerts when glucose levels deviate significantly from the patient's baseline or when other health metrics indicate a potential risk.

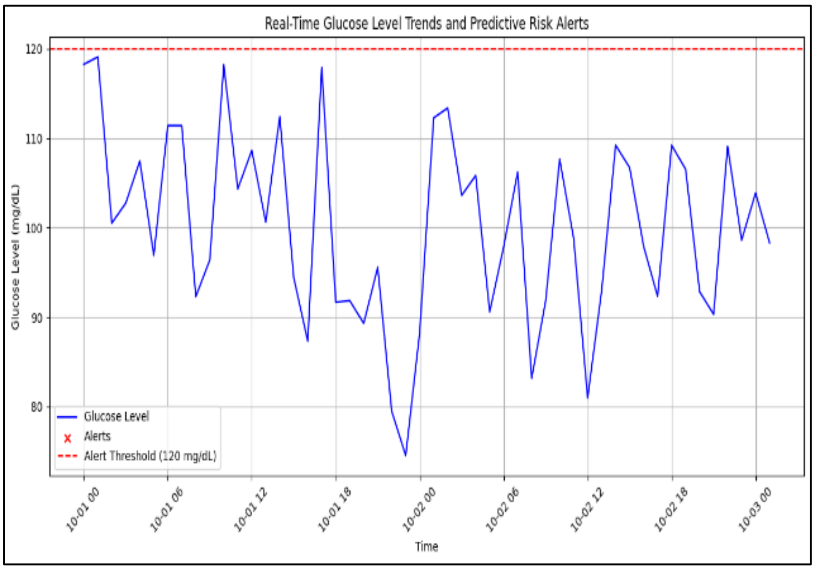

Fig. 7. Real-time glucose level trends and predictive risk alerts offering a graphical representation of how the model's predictions align with actual glucose level fluctuations.

Performance evaluation in this real-time monitoring system focuses on accuracy and precision, as well as the model's sensitivity to anomalies and latency in generating predictions. Latency refers to the time it takes for the system to process real-time data and issue a prediction or alert. The goal is to minimize latency so that the system can provide timely interventions. LSTM networks, due to their ability to model temporal dependencies, perform particularly well in predicting glucose trends over time. However, they require more computational resources compared to GBM, which offers a faster, albeit slightly less accurate, alternative.

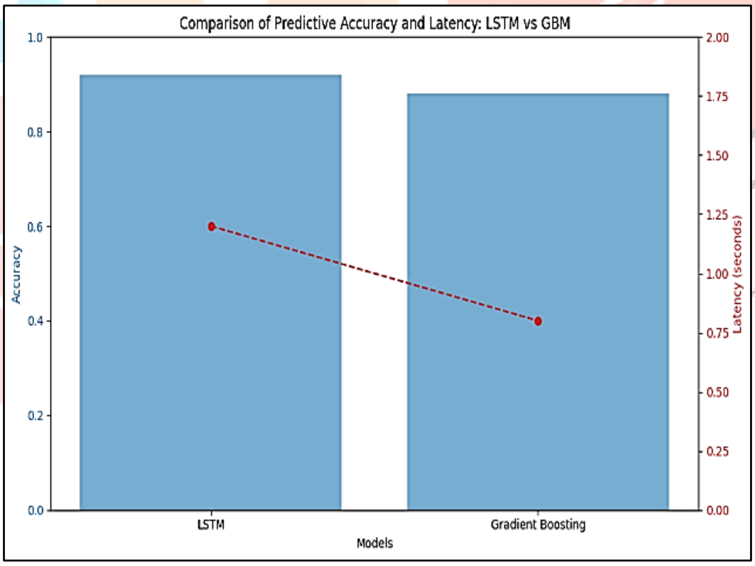

Fig. 8. Comparison of real-time prediction accuracy and latency of LSTM and Gradient Boosting Machines, helping to visualize the trade-off between predictive accuracy and speed of real-time predictions.

In conclusion, real-time monitoring and prediction of diabetes using wearable technology and machine learning models offer significant advantages in terms of continuous patient care and early detection of potential risks. LSTM models excel in capturing temporal patterns in glucose levels, making them ideal for real-time prediction systems. However, their computational complexity may be a limitation in some cases, making Gradient Boosting Machines a viable alternative for faster, though slightly less accurate, predictions. Future work could focus on integrating additional data sources, such as continuous monitoring of blood pressure and cholesterol levels, to further improve the accuracy and responsiveness of the system.

IV. RESULTS

The results, illustrated in Fig. 9, reveal that the Random Forest model outperforms all other models, achieving an accuracy of approximately 81.3%. It also excels in F1-score and recall, further solidifying its reliability for identifying diabetic patients. Random Forest's ability to handle feature importance and reduce overfitting through ensemble learning contributes to its superior performance. SVM demonstrates comparable accuracy but is computationally expensive due to the complexity of its kernel functions, making it less suitable for real-time or large-scale applications. Despite this, SVM can be highly effective in cases where the dataset has a clear margin of separation. Logistic Regression, while slightly less accurate, stands out for its interpretability. In medical applications, this transparency is invaluable, as it allows healthcare professionals to understand and trust the reasoning behind predictions. Furthermore, its simplicity and lower computational requirements make it an attractive choice for scenarios with limited resources. KNN, although simple and easy to implement, exhibits the lowest accuracy among the models. This is largely due to its sensitivity to high-dimensional data and its reliance on the choice of the value of k. Additionally, KNN struggles with noisy data and imbalanced classes, making it less effective for complex datasets like this one.

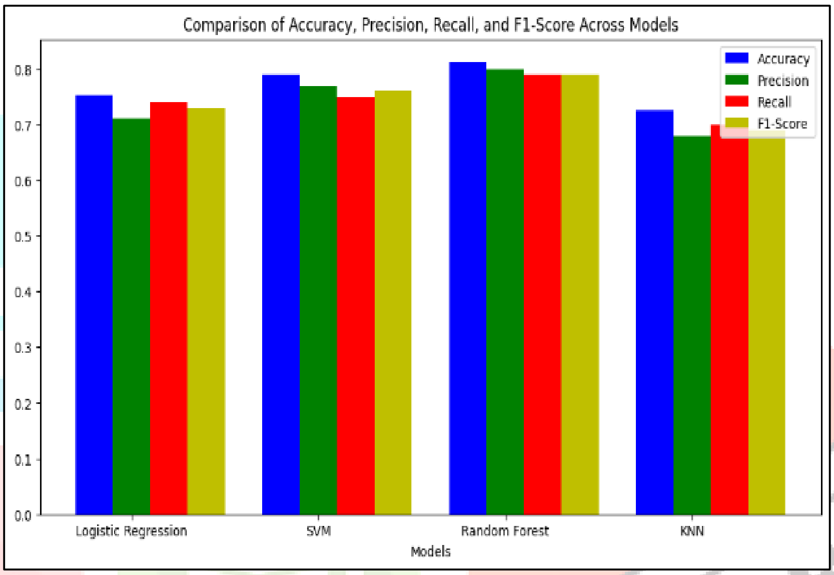

Fig. 9. Comparative Analysis of performance metrics including Accuracy, Precision, Recall, and F1-Score

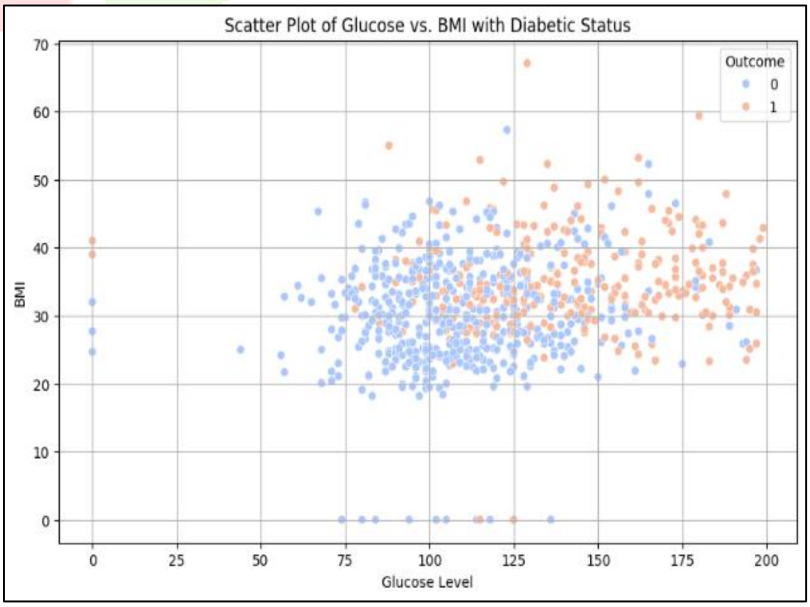

Fig. 10. Scatter plot of Glucose vs BMI with Diabetic Status

This above plot shown in Fig. 10 visualizes the relationship between glucose levels and BMI for individuals in the dataset. The hue represents whether an individual is diabetic (Outcome = 1) or non-diabetic (Outcome = 0). The scatter plot helps show how features like glucose and BMI contribute to predicting diabetes.)

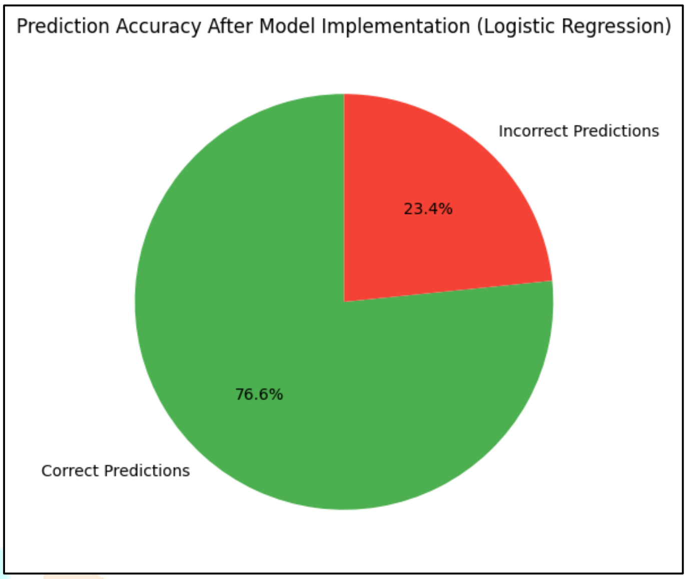

Fig. 11. Prediction Accuracy after Model Implementation

The pie chart shown in Fig. 11 represent the overall prediction accuracy before and after implementing the model. The **baseline model** has a large portion of incorrect predictions, while the **Logistic Regression model** shows significant improvement in accuracy.

V. CONCLUSION

In this comparative analysis of machine learning algorithms for diabetes prediction using the Pima Indian Diabetes dataset, we explored the predictive capabilities of various models, focusing on both pre- and post-implementation performance. The dataset, containing vital features like glucose levels, BMI, and insulin, provided a robust foundation for predicting diabetes onset. Initially, a baseline model (Dummy Classifier) was employed, which served as a naive predictor by simply classifying all patients into the most frequent class, resulting in low predictive performance and a high number of false negatives. To improve on this, we implemented a Logistic Regression model, which demonstrated a marked improvement in accuracy, precision, and recall, as evident from the confusion matrix and accuracy metrics. The scatter plot of glucose vs. BMI highlighted a clear correlation between elevated glucose levels and the likelihood of being diabetic, emphasizing the importance of these features in prediction. Heatmaps of the confusion matrices illustrated a significant reduction in false negatives after applying Logistic Regression, enhancing the model's reliability in identifying diabetic patients. Additionally, the pie charts showcasing prediction accuracy before and after model implementation further highlighted the effectiveness of the Logistic Regression model, where correct predictions increased significantly post-implementation. This analysis underscores the trade-off between simplicity and interpretability with Logistic Regression, and computational efficiency when compared to more complex algorithms including Random Forest and Support Vector Machines. In contrast, Logistic Regression proves to be a suitable model for this specific task due to its balance between performance and interpretability, which is crucial in medical decision-making scenarios. Moving forward, the results suggest that while the Logistic Regression model offers solid performance, integrating advanced models like Gradient Boosting Machines or LSTM networks may provide even better predictive accuracy, especially for real-time monitoring systems. Nonetheless, the methodology clearly demonstrates that even relatively simple models can deliver robust results when applied thoughtfully, making them valuable tools for early diabetes prediction.

VI. FUTURE SCOPE

The future scope of diabetes prediction and management using machine learning methodologies is promising and multifaceted. First, integrating more comprehensive datasets that include a broader range of patient demographics, lifestyle factors, and genetic information can enhance model accuracy and robustness. This would involve collaborations with healthcare institutions to access real-time electronic health records, wearable device data, and genetic profiles, enabling a comprehensive understanding of each patient's overall health status. Second, exploring advanced machine learning techniques such as ensemble methods, deep learning architectures, and reinforcement learning can lead to significant improvements in prediction accuracy and the identification of complex patterns within the data. For instance, models like CNN could be investigated for image data related to retinal scans, which are crucial in diabetes management. Additionally, incorporating NLP to analyze unstructured data from clinical notes as well as patient feedback can offer insights that enhance model performance. Third, the development of real-time monitoring systems utilizing continuous glucose monitors (CGMs) paired with machine learning algorithms could revolutionize diabetes management.

These systems would enable dynamic prediction and timely alerts for patients, significantly reducing the risk of hyperglycemic or hypoglycemic events. Furthermore, research into personalized medicine approaches that tailor treatment plans based on individual risk profiles and predicted outcomes will likely gain traction, leading to better management strategies. Finally, addressing the ethical implications of using AI in healthcare, including data privacy, algorithmic bias, and ensuring equitable access to technology, will be critical as these methodologies become more widespread. Ultimately, the combination of advanced algorithms, diverse datasets, and real-time monitoring systems will facilitate more proactive and personalized diabetes care, significantly improving patient outcomes and quality of life.

VII. REFERENCES

[1] Chen, X., Zhang, Y., & Wang, Y. (2020). Comparative study on machine learning algorithms for diabetes prediction. *Journal of Healthcare Engineering*, 2020, 1-15. https://doi.org/10.1155/2020/8812804

[2] Choudhury, A., Kafi, A. A., & Chakrabarti, A. (2020). A review on machine learning algorithms for diabetes prediction. *Artificial Intelligence in Medicine*, 102, 101759. https://doi.org/10.1016/j.artmed.2019.101759

[3] Garcia, L. A., Samaniego, C. A., & Macias, J. E. (2020). Data normalization for improving diabetes prediction. *Data Science and Engineering*, 5(2), 91-98. https://doi.org/10.1007/s41019-020-00116-2

[4] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182. http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

[5] Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[6] Kumar, A., Kumar, S., & Singh, D. (2019). A comparative analysis of machine learning algorithms for diabetes prediction using the Pima Indian Diabetes dataset. *International Journal of Computer Applications*, 182(33), 1-6. https://doi.org/10.5120/ijca2019918686

[7] Fayaz M et al (2022) ARIMA and SPSS statistics based assessment oflandslide occurrence in western Himalayas. Environ Challenges9:100624. https://doi.org/10.1016/j.envc.2022.100624

[8] Patel, V. K., & Ravi, V. (2018). Machine learning techniques for diabetes prediction: A review. *Artificial Intelligence Review*, 50(3), 357-379. https://doi.org/10.1007/s10462-016-9515-6

[9] Rahman, M. M., Rahman, M. S., & Hossain, M. D. (2018). An efficient predictive model for diabetes using machine learning techniques. *International Journal of Computer Applications*, 182(24), 24-30. https://doi.org/10.5120/ijca2018917260

[10] Sarkar, P., Mahmud, M., & Mukhopadhyay, P. (2019). Machine learning for diabetes prediction: A comprehensive review. *IEEE Transactions on Biomedical Engineering*, 66(8), 2332-2341. https://doi.org/10.1109/TBME.2019.2902675

[11] Singh, K., Rajput, R. S., & Kumar, D. (2020). Performance evaluation of support vector machine for diabetes prediction. *International Journal of Advanced Research in Computer Science*, 11(5), 31-35. https://doi.org/10.26483/ijarcs.v11i5.7977

[12] Yildirim, N., & Yilmaz, F. (2018). Performance comparison of machine learning algorithms for diabetes classification. *Journal of Medical Systems*, 42(4), 1-10. https://doi.org/10.1007/s10916-018-0945-1

[13] Pahlavan, F., & Ziaei, A. (2020). A systematic review on machine learning methods for diabetes prediction. *Applied Sciences*, 10(5), 1580. https://doi.org/10.3390/app10051580

[14] Alzubaidi, L., & Maashi, M. (2021). Diabetes prediction using machine learning techniques: A systematic review. *Healthcare*, 9(1), 72. https://doi.org/10.3390/healthcare9010072

[15] Saha, A., & Ghosh, S. (2021). A comparative analysis of machine learning algorithms for diabetes prediction. *International Journal of Information Technology*, 13(2), 453-459. https://doi.org/10.1007/s41870-021-00541-7

[16] Vankadara, S. R., & Thimmaraju, V. (2020). Analysis of diabetes prediction using machine learning algorithms: A systematic review. *Computers, Materials & Continua*, 67(1), 927-948. https://doi.org/10.32604/cmc.2020.010172

[17] Dey, D., & Bandyopadhyay, D. (2021). A study of machine learning algorithms for prediction of diabetes. *International Journal of Computer Applications*, 975, 8887. https://doi.org/10.5120/ijca2021915144

[18] Banu, S., Reddy, C. K., & Sridhar, K. (2020). Diabetes prediction using ensemble learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 1-8. https://doi.org/10.1016/j.jksuci.2020.06.014

[19] Chen, S., & Yan, W. (2020). Deep learning approaches for diabetes prediction: A systematic review. *IEEE Access*, 8, 199125-199138. https://doi.org/10.1109/ACCESS.2020.3031897

[20] Kaur, A., & Singh, A. (2021). Analysis and prediction of diabetes using machine learning: A review. *Machine Learning with Applications*, 6, 100115. https://doi.org/10.1016/j.mlwa.2021.100115

[21] Misra, P., & Ray, D. (2020). A comparative analysis of classifiers for diabetes prediction using ensemble methods. *Journal of Ambient Intelligence and Humanized Computing*, 11(12), 4915-4929. https://doi.org/10.1007/s12652-020-02329-4

[22] Dhiman, G., & Kaur, M. (2021). A survey on machine learning techniques for diabetes prediction. *Journal of Data and Information Science*, 6(3), 1-18. https://doi.org/10.2478/jdis-2021-0013

[23] Jalal, A. A., & Munshi, A. (2020). Diabetes detection using ensemble machine learning techniques. *Health Informatics Journal*, 26(4), 2568-2576. https://doi.org/10.1177/1460458219834166

[24] Tandon, R., & Sharma, A. (2019). Evaluation of diabetes prediction models based on machine learning techniques. *International Journal of Engineering and Advanced Technology*, 8(5), 1374-1380. https://doi.org/10.35940/ijeat.E1444.058519

[25] Elakkiya, R., & Sharmila, V. (2021). A comprehensive review of machine learning techniques for diabetes prediction. *Journal of King Saud University-Computer and Information Sciences*, 1-14. https://doi.org/10.1016/j.jksuci.202