iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62259

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Advanced Ensemble Framework for Diabetes Outcome Forecasting

Adithya Suresh[1], Ans Benny[2], Minnu Antony[3], Pristy Paul T[4], Rotney Roy Meckamalil[5]
Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***Diabetes is one of the chronic diseases, which is increasing from year to year. Developing an automated system that can detect diabetes patients plays an important role in medical science. A stack classifier model is designed for the detection of diabetes by combining three base estimators such as Random Forest Classifier, LightGBM Classifier, and K-Nearest Neighbors Classifier, and Logistic Regression as meta-classifiers. The data preprocessing includes transforming categorical variables into numerical format. Each base learner is trained on the preprocessed data and predictions are made. In the meta learner stage, Logistic Regression is trained to make predictions based on the predictions of the base learners. The goal of the meta learner is to learn how to combine the predictions of the base learners to make a more accurate final prediction on whether the patient is diabetic or not. The use of a stacking classifier improves prediction accuracy compared to using a single classifier. The developed model gives an accuracy of 98%. The 5-fold cross validation is used to get a more robust estimation of generalization error. Thus, the developed model offers a means to enhance early detection and elevate the quality of care for diabetic patients.***
Index Terms: Diabetes detection, Ensemble model, KNN imputer, healthcare

I. INTRODUCTION

Diabetes is a chronic health condition characterized by high blood sugar levels. It occurs when the body either doesn't produce enough insulin or can't use it effectively. If left unmanaged, diabetes can lead to serious complications such as heart disease, kidney failure, and blindness[1].

Developing a reliable diabetes prediction model is crucial for effective healthcare management. Machine learning techniques, such as RandomForest, KNeighborsClassifier, and LGBMClassifier, play a significant role in predicting diabetes risk[2]. These algorithms analyze factors like age, gender, and various biochemical markers to identify patterns associated with diabetes onset.

Moreover, understanding the significance of early diabetes detection cannot be overstated. Early diagnosis allows for timely interventions and personalized treatment plans, which can significantly improve patient outcomes and reduce the risk of complications associated with diabetes[3]. By leveraging machine learning algorithms to accurately predict diabetes risk, healthcare professionals can provide proactive care and support to individuals at risk, ultimately leading to better healthcare management and improved quality of life.

The increasing prevalence of diabetes globally has underscored the importance of accurate prediction methods. However, existing methods often lack the precision and efficiency required for early detection. As a result, there is a growing need for a reliable system that can predict diabetes early on, aiding in better healthcare management[1].

Manual analysis of healthcare data, with its ever-growing volume and complexity, is becoming impractical. This poses a challenge in identifying individuals at risk of diabetes and implementing timely interventions. Therefore, there is a pressing need for an automated system that can analyze diverse patient data and provide timely predictions.

This automated system would enable healthcare professionals to identify individuals at risk of diabetes more effectively[4]. Additionally, it would empower individuals to take proactive steps towards healthier lifestyles based on personalized risk assessments. By addressing these challenges, our project aims to develop a sophisticated diabetes prediction system using machine learning techniques. This system will fill the existing gap in diabetes prediction methods, contributing to better healthcare management and improved quality of life for individuals at risk of diabetes[2].

II. RELATED WORKS

A. Deep Learning and Specialized Neural Networks

Deep learning architectures, such as Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM) networks, and specialized Deep Neural Networks (DNNs), are gaining traction in healthcare, particularly for tasks like diabetes prediction and detection.

These models excel at learning intricate patterns and relationships in data, making them highly effective in processing large volumes of medical data efficiently. For instance, CNNs are valuable in analyzing medical imaging for conditions like diabetic retinopathy, while BiLSTM networks are adept at capturing temporal dependencies in sequential data from patient records. By incorporating these advanced neural network models into diabetes prediction systems, there is a significant potential for improving accuracy and predictive capabilities, as they enable a deeper understanding of nuanced semantic relationships within medical data, ultimately leading to more informed clinical decisions.[5,6,7]

B. Integration of IoT and Predictive Analytics

Studies propose integrating Internet of Things (IoT) devices with predictive analytics for real-time monitoring of health data, including blood glucose levels. Machine learning techniques like Random Forest, Linear Regression (LR), and Multilayer Perceptron (MLP) are combined to classify diabetes cases and perform predictive analysis. This integration enables continuous monitoring and analysis of health data, leading to proactive diabetes management. Emphasizing the importance of continuous improvement and adaptation in diabetes prediction, suggestions for future enhancements include real-time monitoring, privacy considerations, and multilingual support. These enhancements aim to improve accessibility and functionality, ensuring more effective and personalized diabetes management[8,9,10].

C. Ensemble Learning

Ensemble learning plays a pivotal role in improving predictive accuracy, especially in diabetes prediction scenarios. Techniques like Confusion Matrix Based Integration (CIBM), Soft Voting (SV), AdaBoost, and Random Forest (RF) are deployed to combine the predictions of multiple models, offering more reliable insights into diabetes onset and classification[11,12]. Understanding ensemble learning principles such as bagging, boosting, and stacking is crucial for optimizing accuracy. Bagging reduces variance by aggregating independent model predictions, while boosting iteratively corrects errors to enhance precision. Stacking leverages metalearners to fuse model outputs, maximizing predictive power. In practical terms, implementing Random Forest involves finetuning hyperparameters like tree count and depth for optimal performance. Additionally, employing feature selection techniques and rigorous validation on independent datasets ensures model robustness and guards against overfitting, making Random Forest a potent tool in real-world diabetes prediction tasks[3,8].

D. Feature Selection and Engineering

Feature selection is indeed a crucial aspect of building accurate predictive models, particularly in domains like diabetes prediction. Techniques such as analysis of variance (ANOVA), Chi-square, and Principal Component Analysis (PCA)[12,11] are commonly used to identify the most informative features for prediction. Strategies like synthetic data generation and oversampling minority classes can also improve model performance by ensuring a balanced representation of all classes. In feature selection and engineering for diabetes prediction, a comprehensive approach integrates various methods. This includes using filter methods like correlation analysis, wrapper methods such as forward/backward selection, and embedded methods like LASSO regularization to prioritize relevant features[13]. Feature engineering techniques are equally important, involving tasks like feature scaling, dimensionality reduction via PCA, and creating new features that capture unique data patterns. Ongoing research also focuses on developing novel feature selection algorithms tailored to medical datasets, addressing challenges such as imbalanced class distributions and high-dimensional feature spaces. These advancements ensure that selected features are not only relevant but also robust and reflective of underlying data patterns, thereby enhancing the accuracy and reliability of predictive models in healthcare applications like diabetes prediction.

E. Automated Pipelines and Model Optimization

Automated pipelines are fundamental in optimizing machine learning models for accurate diabetes prediction. These pipelines streamline model development by integrating crucial techniques like feature selection, missing value imputation, and cross-validation.[14] Feature selection ensures that only relevant variables are considered, reducing noise and enhancing model performance. Imputation methods fill in missing data, ensuring a more complete dataset for analysis. Crossvalidation validates the model across different data subsets, ensuring its robustness[15]. Model optimization is equally crucial, involving parameter fine-tuning to achieve optimal predictive accuracy. Techniques like grid search and randomized search are employed to explore parameter spaces effectively. Automating these processes in the pipeline not only saves time and resources but also maximizes the predictive capabilities of diabetes prediction models.

F. Evaluation Metrics and Transfer Learning

Evaluation metrics and transfer learning are instrumental techniques for improving model performance, especially in scenarios with imbalanced datasets and limited training data. Transfer learning facilitates seamless adaptation across diverse datasets, bolstering model generalization[16]. Simultaneously, data augmentation enriches training sets with synthetic data, fortifying model robustness. Key metrics like accuracy, precision, recall, F1-score, and ROC-AUC play crucial roles in assessing model performance, offering insights into classification accuracy and error minimization. Aligning metric selection with study objectives ensures a thorough evaluation, highlighting the model's effectiveness in addressing specific challenges. This strategic combination elevates model performance across diverse domains and tasks[17].

G. Applications and Case Studies

Real-world applications and case studies offer valuable insights into the practical implications of research findings. For instance, in developing diabetes prediction systems, there's a focus on technological aspects like Python Django and machine learning algorithms, highlighting the iterative development process. The goal is to improve patient care through early detection and intervention while adapting to emerging trends in diabetes research and diagnostics. Key components such as feasibility studies, system analysis, and continuous improvement are emphasized to streamline product development and ensure sustained relevance. Future enhancements include fine-tuning analysis methods, incorporating contextual understanding, leveraging transfer learning techniques, and exploring customization features for patients. These examples showcase how technology and methodologies are effectively applied to enhance patient care and predictive systems[16].

III. MATERIALS AND METHODS

In this study, we developed a diabetes prediction model using machine learning techniques. Our approach involved preprocessing the dataset to prepare it for training the model and then evaluating its performance.

A. Diabetes Dataset

Data has been extracted from a diabetes dataset stored in "Diabetesmendeley.csv"[18]. The dataset contains information on several factors such as age, gender, blood test results (e.g., glucose levels, cholesterol), and other health indicators. The dataset contains a wide range of medical information and detailed laboratory analysis attributes gathered from a diverse group of 826 individuals. This group comprises 128 individuals with diabetes, 96 without diabetes, and 40 subjects that are expected to develop diabetes.

B. Data Preprocesing

Data preprocessing plays a crucial role in preparing our dataset for training the diabetes prediction model. Initially, we split the dataset into features (X) and the target variable (y), excluding irrelevant columns such as 'ID' and 'NoPation'. Subsequently, the data was divided into training and testing sets with a 70:30 ratio to ensure robust model evaluation. To handle missing values, various techniques such as imputation or deletion were employed, ensuring the integrity of the dataset. Additionally, the target variable 'CLASS' was transformed into a binary format, where 'N' (non-diabetic) is encoded as 0 and 'Y' (diabetic) as 1, while gender was converted into a numerical format, with 'M' (male) represented as 1 and 'F' (female) as 0. Missing values were addressed through techniques like mean or predictive imputation, while numerical features were normalized to scale them to a similar range, avoiding dominance. Categorical variables were encoded into numerical format, ensuring compatibility with machine learning algorithms. Pertinent features extracted from the preprocessed medical data included demographic attributes, clinical measurements such as glucose levels, blood pressure, and medical history indicators. Additionally, features like family history of diabetes, lifestyle factors, and medication adherence were considered to enhance predictive accuracy. Advanced techniques like word embeddings were explored to capture nuanced semantic relationships within medical records and diagnostic information, thereby enriching the predictive capabilities of the model[15].

C. Ensemble Model Classifiers

The project utilizes a Stacking Classifier for ensemble modeling, incorporating base learners like LightGBM, Random Forest, and K-Nearest Neighbors (KNN), along with a Logistic Regression meta-learner. This ensemble approach combines the strengths of diverse models to enhance accuracy and robustness in diabetes prediction[15].

1) *Random Forest:* Random Forest is a classifier that improves predictive accuracy by combining multiple decision trees trained on different subsets of the dataset. It uses majority voting to predict the final output, preventing overfitting and ensuring efficiency even with large datasets. The process involves randomly selecting subsets of data points and features, building decision trees on these subsets, and combining predictions through majority voting or averaging for classification or regression tasks[9].
2) *LightGBM:* LightGBM, an open-source gradient boosting framework developed by Microsoft, is known for its efficiency, scalability, and accuracy. It optimizes memory usage and training time through techniques like selective instance retention during training and employs a leaf-wise tree growth strategy. Additionally, it utilizes Gradient-Based One-Sided Sampling (GOSS) to prioritize important data instances, speeding up training without sacrificing accuracy. Overall, LightGBM provides quick and effective executions for machine learning tasks, handling overfitting well even with smaller datasets[19].
3) *K-Nearest Neighbors:* K-Nearest Neighbour (K-NN) is a versatile Supervised Learning algorithm for Classification tasks, relying on data point similarities measured through metrics like Euclidean distance. It adapts dynamically to local data patterns by finding the K nearest neighbors and making predictions based on their majority vote or average values. The KNN imputer in sci-kit-learn showcases this adaptability by using the Euclidean distance matrix to impute missing values, prioritizing non-missing coordinates for accurate estimation, and employing a weighted distance calculation for improved imputation quality[20].
4) *Logistic Regression:* Logistic Regression (LR) is a supervised machine learning algorithm used for binary classification tasks. It predicts the probability that an instance belongs to a given class using a sigmoid function, which transforms inputs into probabilities ranging from 0 to 1. LR optimizes parameters during training to maximize the likelihood of observed class labels and applies a decision threshold (usually 0.5) for classification. It is simple, interpretable, and wellsuited for scenarios requiring binary predictions[21].

D. Model Training

The project adopts a systematic approach known as the Stacking Classifier ensemble model, which integrates the predictive capabilities of multiple base models—specifically, the Random Forest Classifier, LightGBM Classifier, and KNearest Neighbors Classifier—with a Logistic Regression meta-learner. This method involves a structured process of training and prediction.During the training phase, the dataset is split into a training and validation set. The training set is further divided into 5 folds for cross-validation. Each base model is then trained on 4 folds of the training data, while predictions are made on the 5th fold[15]. This ensures robustness and helps prevent overfitting by using different subsets of data for training and validation. The Random Forest Classifier constructs decision trees based on random subsets of the training data, with each tree independently predicting the likelihood of diabetes[9]. The LightGBM Classifier employs gradient boosting techniques to iteratively enhance predictive performance by minimizing residual errors from previous iterations. Similarly, the K-Nearest Neighbors Classifier classifies instances based on the majority class of their nearest neighbors in the feature space[19,20]. In the meta-learner stage, the predictions from the trained base estimators are combined to produce a more accurate final outcome forecast. Predictions from the base learners are used as input for the metamodel. This involves obtaining predictions from each base estimator for a given patient dataset, feeding these predictions as features into the meta-learner alongside the actual class labels, and training the Logistic Regression model on this augmented dataset to learn the optimal combination of base learner predictions[21]. Finally, utilizing the trained metalearner, final predictions are generated for new, unseen patient data. The meta-learner adjusts the final prediction by weighing the predictions of the base estimators based on their individual performance, thereby leveraging the collective insights of the ensemble framework[22]. This systematic approach ensures superior accuracy in diabetes outcome forecasting, providing valuable support for early detection and intervention strategies in diabetic patient care.

E. Evaluation Metrics

To assess the model's effectiveness in predicting diabetes, the final stage involves evaluating its performance on the testing dataset using key metrics such as accuracy, precision, recall, and F-score. This comprehensive evaluation provides insights into the model's predictive capability. To enhance interpretability, the model generates probability outputs for the likelihood of diabetes (P(Diabetes)) and the probability of being normal (P(Normal)) for each individual classifier KNN, RF, and LGBM. These probabilities are then averaged across the classifiers to derive a consolidated probability score for each outcome, ensuring a consistent prediction approach. Following this, the prediction outcome (result1) is determined as "Positive" if the model predicts diabetes (1) and "Negative" if no diabetes is predicted (0).

The classification helps with informed decision-making and risk management strategies based on health indicators. We also conducted a thorough analysis to identify the most important features that significantly impact the prediction of diabetes. This analysis assists healthcare professionals in prioritizing risk factors and developing specific interventions for the early detection and effective management of diabetes. By following this systematic approach for evaluation and analysis, the model provides actionable insights into diabetes prognosis. This helps healthcare practitioners implement personalized care strategies, ultimately improving patient outcomes.

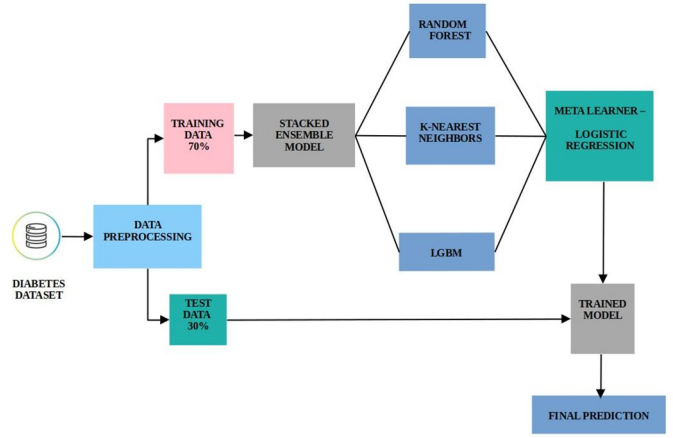


Fig. 1. Architecture Diagram of the proposed model

IV. ALGORITHM

A. *Algorithm for Diabetes Prediction*

1) Imports: Import required libraries such as pandas, scikitlearn modules, and lightgbm.
2) Define Views: Define views for the Django web application, including ”home”, ”predict”, and ”result”.
3) Data Preprocessing:
   - Load the dataset ”Diabetes _mendeley.csv” into a pandas DataFrame.
   - Convert the target variable ”CLASS” to binary encoding (0 for ’N’, 1 for ’Y’).
   - Encode the ”Gender” column into numerical format (1 for ’M’, 0 for ’F’).
   - Split the dataset into features (X) and target (y) variables.
4) Split Data: Split the dataset into training and testing sets using train test split() method from scikit-learn.
5) Initialize Models: Initialize base models (Random Forest, LightGBM, K-Nearest Neighbors) and a metaclassifier (Logistic Regression).
6) Initialize Stacking Classifier:
   - Initialize a StackingClassifier with the base models and the meta-classifier.
   - Set parameters like estimators and cross-validation folds for the meta-classifier.
7) Model Training: Fit the StackingClassifier on the training data (X train, y train).
8) Data Collection and Processing: Retrieve input values (Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI) from the user via request.GET.
9) Preprocess Input: Convert Gender input to numerical format (1 for ’M’, 0 for ’F’).
10) Make Prediction:
   - Use the trained StackingClassifier (stack clf) to predict the outcome based on user input.
   - If the prediction is 0, classify the outcome as ”Negative”; if 1, classify as ”Positive”.
11) Model Evaluation:
   - Predict outcomes on the test set (X _test) to evaluate model accuracy using accuracy score().
   - Print the accuracy score of the model on the test set.
12) Render Result: Render the prediction outcome (”result1”) and any additional information on the ”predict.html” template.

13) End of Algorithm

This algorithm outlines the process of building a diabetes outcome forecasting system using a Stacking Classifier in a Django web application.

V. RESULT

Python 3.10 was utilized to develop the machine learning models in this study. These models were implemented and executed within a Jupyter Notebook environment using the sci-kit Learn library. The experiments were conducted on a system running a 32-bit Windows 11 operating system, equipped with a Ryzen 5 Quad Core CPU. The CPU had a Base Frequency of 2.10 GHz and a Max Turbo Boost Frequency of up to 3.7 GHz. With an impressive accuracy of 98%, the developed model outperformed single classifiers, providing a robust solution for early detection of diabetes. This success underscores the effectiveness of combining multiple base classifiers with a meta-classifier to accurately identify diabetes patients. Moreover, the integration of the Pythonbased backend model with a user-friendly frontend, developed using HTML, CSS, and Django, created an intuitive interface for healthcare professionals to interact with the system. This frontend design ensured accessibility and ease of use, enhancing the overall usability of the automated diabetes detection system.

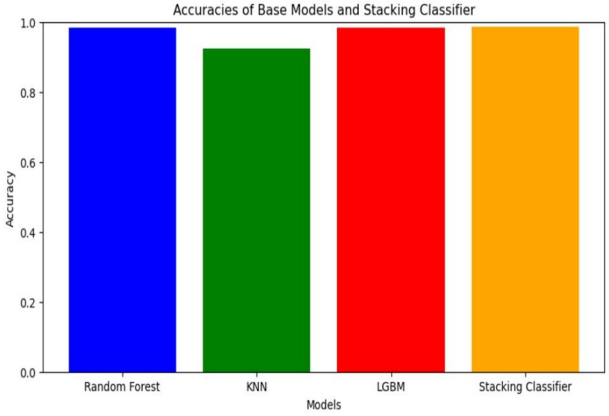


Fig. 2. Comparison of accuracies

In addition to high accuracy, the model's performance has been verified through rigorous testing on different datasets, further confirming its reliability in real-world scenarios. Advanced learning techniques such as feature selection and hyperparameter tuning have been employed to enhance the

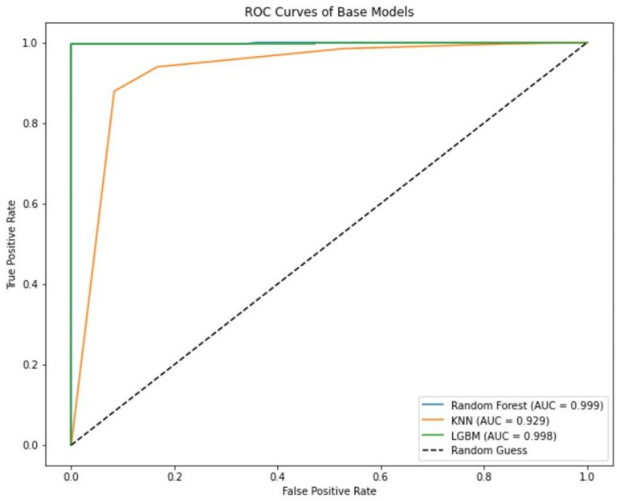


Fig. 3. ROC Curve for base classifiers

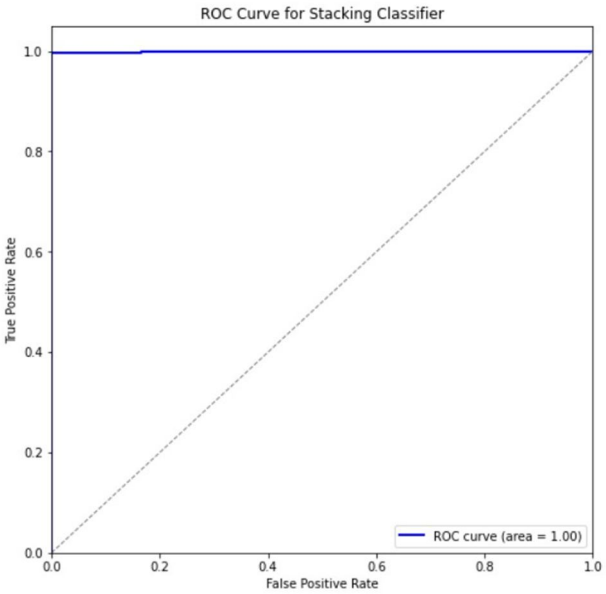


Fig. 4. ROC Curve for Stacking Classifier

robustness and generality of the model[23]. Regular updates and improvements to the system ensure alignment with medical standards and technological advancements, thereby enhancing its performance and serving as a vital tool for early disease detection and health management. Receiver Operating Characteristics (ROC) curves and the Area Under the ROC Curve (AUC) were utilized to assess the predictive performance of the model and determine the effectiveness of its predictions[3].

VI. FUTURE SCOPE

The future scope and developments of diabetes prediction systems hold significant promise for advancing healthcare. Several areas can be explored to enhance these systems.

1) Fine-tuning analysis methods can be pursued, enabling the identification and categorization of different risk factors and predictive features related to diabetes. This provides a detailed understanding, allowing for accurate prediction and targeted interventions.

2) Incorporating contextual understanding is crucial. By leveraging contextual data such as lifestyle factors, genetic information, and medical history, prediction systems can better grasp the complex interactions contributing to diabetes development. This enables a more comprehensive assessment tailored to individual patients.

3) Transfer learning techniques can be employed to leverage existing medical knowledge and adapt it to new datasets. Fine-tuning pre-trained models on diverse patient populations and medical contexts enhances system performance and generalization capabilities.

4) Customization features for patients are also worth exploring. Allowing input of personal health data, preferences, and goals for diabetes management can personalize the system to align with unique circumstances and needs.

5) Privacy and ethics considerations are essential. Compliance with healthcare regulations and implementing privacy safeguards ensure responsible handling of patient data.

6) Support for multilingual and multicultural populations is another area of growth. Adapting systems to different languages, cultural contexts, and healthcare systems expands their applicability and accessibility.

7) Incorporating feedback mechanisms from healthcare providers and patients can improve system accuracy and usability. This collaborative approach enhances patient engagement and encourages proactive management of diabetes.

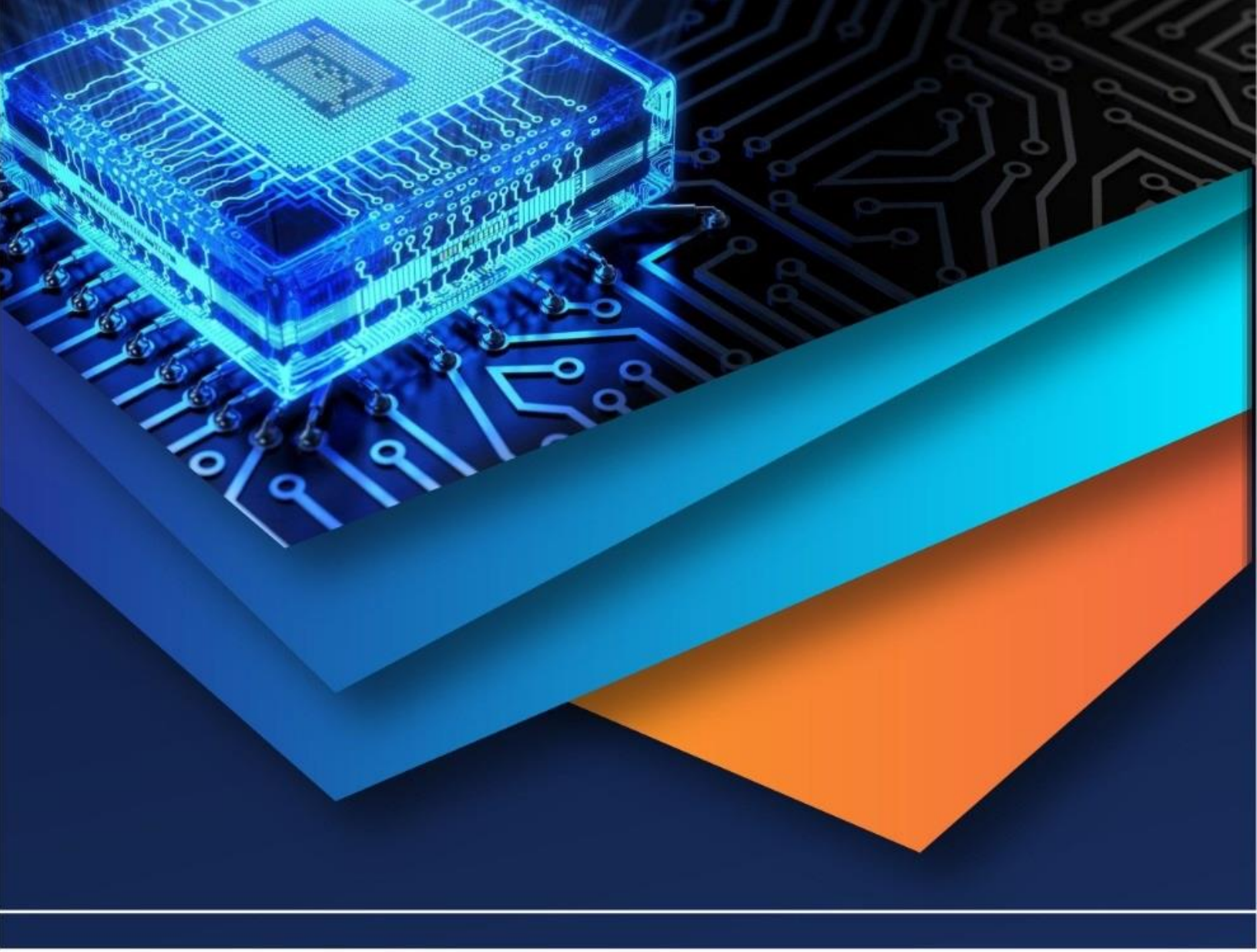
In summary, the future scope of diabetes prediction systems is vast. Advancements in analysis, contextual understanding, transfer learning, customization, privacy and ethics, multilingual support, and collaborative feedback will contribute to more effective and patient-centered systems, ultimately improving diabetes management and healthcare outcomes.

VII. CONCLUSION

In conclusion, the diabetes prediction model represents a significant advancement in addressing the need for early diabetes detection amidst rising diabetes rates. A machine learning architecture integrating data preprocessing, feature engineering, and ensemble modeling was successfully developed to predict diabetes status. The project underscores the effectiveness of ensemble techniques, such as the StackingClassifier, in combining multiple base models to achieve higher accuracy. This project illustrates the potential of machine learning in analyzing and predicting diabetes status based on health indicators. The ability to capture complex patterns using Random Forest, KNN, and LightGBM models within a StackingClassifier framework was demonstrated. Utilizing Logistic Regression as a meta-classifier enhances both interpretability and generalization. Moreover, the project emphasizes the importance of continuous improvement and adaptation in diabetes prediction. Future enhancements could include incorporating genetic or lifestyle factors for a more comprehensive analysis. Advanced techniques like transfer learning and customization could further aid in contextual understanding and personalized risk assessment. By building on these insights, a solid foundation for future advancements in diabetes detection has been laid. The focus on model training, evaluation, and system integration ensures a robust and scalable solution. Future directions may explore real-time monitoring, privacy considerations, and multilingual support to improve accessibility and user experience.

REFERENCES

[1] Diabetes Gojka. (Jul. 2019). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023.
[2] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, “A Machine Learning Approach to Predicting Diabetes Complications,” Healthcare, vol. 9, no. 12, Art. no. 12, Dec. 2021.
[3] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, “An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study,” Sensors, vol. 22, no. 14, p. 5247, Jul. 2022.
[4] Sugandh, Fnu et al. “Advances in the Management of Diabetes Mellitus: A Focus on Personalized Medicine.” Cureus vol. 15,8 e43697. 18 Aug. 2023,
[5] WHO. (Apr. 2023). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023
[6] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh,
A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, “An optimization-based diabetes prediction model using CNN and bidirectional LSTM in real-time environment,” Appl. Sci., vol. 12, no. 8, p. 3989, Apr. 2022.
[7] K. Kannadasan, D. R. Edla, and V. Kuppili, “Type 2 diabetes data classification using stacked auto encoders in deep neural networks,” Clin. Epidemiol. Global Health, vol. 7, no. 4, pp. 530–535, Dec. 2019.
[8] U. Tariq, I. Ahmed, A. K. Bashir, and K. Shaukat, “A critical cybersecurity analysis and future research directions for the Internet of Things: A comprehensive review,” Sensors, vol. 23, no. 8, p. 4117, Apr. 2023
[9] S. Saranya and S. Bobby, “COVID-19 patient health prediction using boosted random forest algorithm,” Data Anal. Artif. Intell., vol. 3, no. 2, pp. 64–68, Feb. 2023.
[10] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, “Machine learning based diabetes classification and prediction for healthcare applications,” J. Healthcare Eng., vol. 2021, pp. 1–17, Sep. 2021.
[11] H. M. Debernehand I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” Int. J. Environ. Res. Public Health, vol. 18, no. 6, p. 3317, Mar. 2021.
[12] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, “Chi-square and PCA based feature selection for diabetes detection with ensemble classifier,” Intell. Autom. Soft Comput., vol. 36, no. 2, pp. 1931–1949, 2023.
[13] Lixin Cui, Lu Bai, Yue Wang, Philip S. Yu, Edwin R. Hancock,”Fused lasso for feature selection using structural information”,Pattern Recognition, Volume 119,2021, 108058
[14] Abnoosian, K., Farnoosh, R., and Behzadi, M. H. (2023). ”Prediction of diabetes disease using an ensemble of machine learning multi-classifier models”, BMC Bioinformatics, 24(1), 337,1471-2105,2023
[15] K. Alnowaiser, ”Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model,” in IEEE Access, vol. 12, pp. 16783-16793, 2024.
[16] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, “Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients,” NPJ Digit. Med., vol. 4, no. 1, p. 109, Jul. 2021.
[17] Alhassan Mumuni, Fuseini Mumuni,”Data augmentation: A comprehensive survey of modern approaches”, Array, Volume 16, 2022, 100258
[18] Sahid, Abdus ; Ul Hoque Babar, Mozaddid; Uddin, Md Palash (2024), “Multiclass Diabetes Dataset”, Mendeley Data, V1, https://data.mendeley.com/datasets/wj9rwkp9c2/1.
[19] B. M. Kanber, A. A. Smadi, N. F. Noaman, B. Liu, S. Gou and M. K. Alsmadi, ”LightGBM: A Leading Force in Breast Cancer Diagnosis Through Machine Learning and Image Processing,” in IEEE Access, vol. 12, pp. 39811-39832, 2024.
[20] A.Juna,M.Umer,S.Sadiq,H.Karamti,A.A.Eshmawi,A.Mohamed,and I. Ashraf, “Water quality prediction using KNN imputer and multilayer perceptron,” Water, vol. 14, no. 17, p. 2592, Aug. 2022.
[21] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, Mar. 2002.
[22] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, “Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques,” in Intelligent and Cloud Computing, Singapore, 2021, pp. 399–409.
[23] Vincent, A. M., Jidesh, P. (2023). An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms. Scientific Reports, 13(1), 4737. https://doi.org/10.1038/s41598023-32027-3



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY