International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Feature-Optimized and Explainable Machine Learning Framework for Early Diabetes Prediction Using Hybrid Clinical and Lifestyle Data

Sandesh Dinkar Mhaske[1], Dr. Namrata Gadgil[2]

[1]Computer Application Department of Computer Science, JSPM University, Pune, Maharashtra, India
Email: *sandeshmhaske2362[at]gmail.com*

[2]Associate Professor Department of Computer Science, JSPM University, Pune, Maharashtra, India
Email: *nsg.scos[at]jspmuni.ac.in*

Abstract: ***However, despite its increasing significance, diagnosis of this disease often happens at a relatively advanced stage and requires effective and timely treatment of associated problems. Therefore, the present study deals with the design of a machine learning-based system for prediction of the disease before any complications take place and involves analysis of patients' clinical and behavioral data. For the purposes of analysis, a database with patient information concerning such parameters as age, gender, BMI, hypertension, heart disease, HbA1c and blood glucose level will be utilized. As a part of preprocessing, missing values will be replaced, categorical features will be converted into numerical and additional actions will be taken in order to improve the quality of collected data. Furthermore, in order to eliminate redundant features, the selection process will be performed prior to application of the models under study. In the course of experimentation, Logistic Regression, Random Forest and XGBoost algorithms will be analyzed. Based on the obtained results, the following accuracy scores will be achieved: 86% for Logistic Regression, 89% for Random Forest and 92% for XGBoost. From the perspective of the experiment outcomes, it can be concluded that ensemble approaches significantly outperform standard techniques thanks to better patterns capturing capabilities. To improve the interpretability of the predictions, methods like SHAP, which belong to the field of explainable AI, have been used as well. With the help of SHAP, it becomes possible to determine the role played by each variable in making the predictions. It has been found that blood glucose level, HbA1c level, body mass index (BMI), and age play the most important roles in predicting diabetes. On the whole, the suggested methodology has successfully achieved a balance between predictive power and interpretability of the model.***

Keywords: Machine Learning, Feature Selection, Predictive Modelling, Health Informatics, Classification Algorithm, Data Preprocessing, Clinical Decision Support Systems, Risk Assessment

1. Introduction

Diabetes is a chronic illness characterized by blood sugar metabolism and has become a global public health challenge because of its fast rate of transmission and severe complications. Recent medical reports show that the incidence of diabetes has continued to increase, posing a great strain on health care services [18]. Early detection of patients is important in the prevention and treatment process since the disease may cause various complications including heart attack, kidney problems, and blindness. Nonetheless, despite improvements in modern medicine, early diagnosis of diabetes patients is still a difficult undertaking.

Conventionally, laboratory tests have been used in the detection of diabetes, but they have limitations regarding accessibility and costs. Machine learning approaches have been adopted extensively in the medical field to enhance predictive models [1] [2]. Several experiments have been carried out using algorithms like decision trees, support vector machines, and random forest for diabetic prediction, recording satisfactory results [6]. Nonetheless, most algorithms are inefficient when it comes to handling redundancy and irrelevancy.

A critical gap in the existing literature involves the inadequacy of an efficient feature optimization approach for enhancing model accuracy and simplicity. Most of the models take advantage of all the features without selecting the relevant features, making the models less efficient and complex.

To solve this problem, in this research, an ML model for the optimal feature selection process that enables early prediction of diabetes is developed. In this research, the aim is to identify the most influential features in relation to diabetes and design an efficient model for prediction. With the use of feature selection algorithms combined with classification algorithms including logistic regression, random forest, and XGBoost [9], accuracies of 86%, 89%, and 92%, respectively, are achieved.

Moreover, SHAP, which is part of the explainable AI approaches, is used to optimize model predictions through identification of the contribution of each feature to predictions.

2. Problem Statement

In many instances, diabetes is detected when symptoms have worsened, as most methods for diagnosing diabetes depend more on laboratory investigations and medical records than anything else. Even though this approach works well, it is not always applicable to detecting early signs of diabetes. Besides, most conventional strategies emphasize clinical measures while disregarding life choices, which may contribute to the emergence of diabetes. Besides, many existing algorithms for predicting diabetes emphasize the

importance of high prediction accuracy while being relatively hard to interpret [13]. As a result, the process by which conclusions are drawn from the data is not well understood, thus making it challenging to use the algorithm. Moreover, many data sets have a lot of unnecessary attributes, which can adversely affect the results of machine learning processes. Hence, there is a need for an approach that will predict the likelihood of developing diabetes through an integrated evaluation of clinical measures and life choices at an early stage. The method should be capable of generating accurate predictions while providing explanations regarding the rationale behind its decisions.

3. Literature Review

Today, the spread of diabetes has reached alarming proportions around the globe [18]. Therefore, various means to diagnose this condition early on have been developed by scientists. Historically, diagnosis of diabetes has been carried out by using laboratory testing as well as clinical examination. Even though these approaches have proven effective in practice, they cannot always be applied in the process of predicting this disease at the early stages of its development.

To enhance the accuracy of the prediction models, feature selection methods are typically used in machine learning research projects [17]. Such approaches as RFE (Recursive Feature Elimination), PCA (Principal Component Analysis), and correlation-based algorithms are employed for finding the best features for building a predictive model as well as for getting rid of irrelevant information [4].

Another significant aspect explored in past literature reviews is the comparison between various machine learning algorithms. It was observed that ensemble techniques could provide greater accuracy than single models, but they were usually complicated and needed higher computational power. At the same time, Logistic Regression proved to be accurate enough if used in conjunction with suitable data preprocessing and feature selection techniques. Thus, selecting a machine learning algorithm depends on the type of dataset and the task assigned to the computer systems [1].

Machine Learning algorithms are increasingly being used for clinical decision support systems. The systems use patient data including age, BMI, level of glucose, and any diseases suffered by the patient in order to be able to predict how patients will behave [3].

But despite all this development, there are also a few constraints that cannot be overlooked. For example, some algorithms become too complex for implementation, while other algorithms can easily be implemented but may not be very accurate. It should also be mentioned that algorithms developed through one data set can fail when applied to another data set.

In this research paper, the objective will be to develop a machine-learning based algorithm that allows for the early prediction of diabetes, utilizing optimal feature selection. The emphasis will be put on improving the accuracy of the predictions, while at the same time minimizing the complexity of the algorithm.

4. Objectives of Research

Objectives of the study include:

1) To analyze the data and learn more about the information regarding patients such as their age, gender, BMI, hypertensive problems, heart disease, level of HbA1C and Blood Glucose level.
2) To perform data cleaning through methods such as handling of missing values, removing unnecessary variables and transforming the data from categorical to numeric format.
3) To filter out those features which contribute towards the predictive model and eliminate less valuable ones.
4) To implement several Machine Learning models such as Logistic Regression, Random Forest and XGBoost in diabetes prediction.
5) To compare the performance of these models using accuracy and other evaluation methods.
6) To use SHAP technique to understand how different features affect the prediction results.
7) To show the results using graphs like confusion matrix, ROC curve, and feature importance.
8) To find out the best model (XGBoost) for predicting diabetes.

5. Related Work

The growing prevalence of Diabetes has resulted in increased scientific interest in building efficient and precise prediction models by using machine learning methodologies [3] [5]. Traditionally, the disease detection process is based on medical testing and analysis. These methods, however, do not guarantee its early detection. In turn, the increasing number of data-driven algorithms can significantly enhance the process of prediction and help make informed decisions in healthcare. Machine learning models allow processing huge amounts of medical data and finding underlying patterns. Such machine learning algorithms as Logistic Regression, Random Forest, and XGBoost have been successfully applied to build predictive models of diabetes. According to existing literature, the ensemble approach is the most promising for achieving high accuracy because it can take into account complex interdependencies between features. Nevertheless, ensembles are harder to interpret and computationally inefficient. The second choice should be Logistic Regression, even though it yields less accurate results. This model can be very helpful provided that appropriate data preprocessing and feature selection procedures have been applied. Another topic that attracted the attention of many researchers is feature selection from health care datasets. The methods of recursive feature elimination, principal component analysis, and correlations-based approaches are often applied to eliminate redundant and irrelevant features in order to increase model efficiency [17]. Along with feature selection, data pre-processing and balancing procedures can be also considered important tasks that can lead to a better classification performance. Specifically, data imputation to fill missing data gaps, the application of categorical encoding to transform categories into numeric values, and solving the problem of class imbalance with such methods as SMOTE, proved to

contribute to improving the results achieved during classification. More recently, researches aimed at applying machine learning algorithms to developing decision support systems for clinical practice emerged. Namely, patient-related data such as age, body mass index, blood glucose and HbA1c levels, and medical history become useful when it comes to diagnosing certain conditions and developing personalized treatment plans. In addition, the emergence of explainable AI tools such as SHAP was suggested in order to enhance the interpretability of ML results [12]. However, there are some drawbacks. In particular, many models are complicated to understand in healthcare contexts. Furthermore, there is a lack of attention towards designing systems which consider all three metrics, i.e., performance, interpretability, and efficiency. Finally, even the most efficient algorithms are prone to generalization problems since they might be trained on one data set only. This research suggests creating a feature-optimized machine learning framework for detecting diabetes with the help of Logistic Regression, Random Forest, and XGBoost approaches. Besides high performance, this framework will have a clear explanation due to SHAP methodology.

6. Research Gap

Despite many advances in machine learning application for Diabetes prediction, many research limitations still remain unaddressed in this area. While there have been a lot of studies dedicated to enhancing the accuracy of predictions, very few compare various algorithms to see how well they predict early Diabetes development. It is thus very difficult to establish what method could be used for the optimal results.

Another challenge related to Diabetes prediction algorithms is inconsistency in terms of the used datasets, as well as feature selection process and performance evaluation criteria employed by researchers. Various data sets and assessment methods make it hard to evaluate all results on the same scale and compare them. Another limitation is that in certain cases, not fully adequate data is used to conduct research. In such instances, the results obtained may not be very accurate.

Finally, there is a lack of focus on the data preparation process, where the data undergoes initial processing and is ready to be processed in further analysis. The processes of handling missing values, encoding, normalization, and selecting features can be erroneous, which affects prediction performance adversely.

Moreover, one more significant gap is associated with model explainability. Most of the studies are concentrated on models which offer high accuracy, but fail to give a detailed explanation of their working process [13]. This absence of clarity results in lower confidence of healthcare practitioners in the system and makes it harder to implement such models.

Apart from that, the problem with implementation can be observed as well since all the studies conducted up till now are mostly experimental. The absence of practical solutions based on which it would be possible to implement the system in healthcare facilities and detect diabetes at an early stage can be considered as one of the gaps of research in this area.

Accordingly, one of the possible ways of overcoming these gaps includes designing a uniform machine learning algorithm which utilizes proper data pre-processing, feature selection, and analysis of the efficiency of models such as Logistic Regression, Random Forest, and XGBoost. Applying SHAP methods can make models more understandable [12].

7. Proposed System Architecture

This system uses a step-by-step approach to predict early diabetes. FFirstly, the patients' data are collected and then prepared by removing any errors, inconsistencies, and missing data from the patients' records.

After preprocessing patients' data, certain features are chosen to enhance the performance of the model without making it complicated. Next, different machine learning models like Logistic Regression, Random Forest, and XGBoost are used for training and testing.

Performance metrics are used to determine how well these models perform. It has been shown that the ensemble models provide better prediction results. To make the models interpretable, SHAP values can be used to assess the contributions of each feature in the prediction process. Finally, the system produces an output which classifies the patient's health condition. Is he diabetic or not? See Fig. 1 for the system architecture.

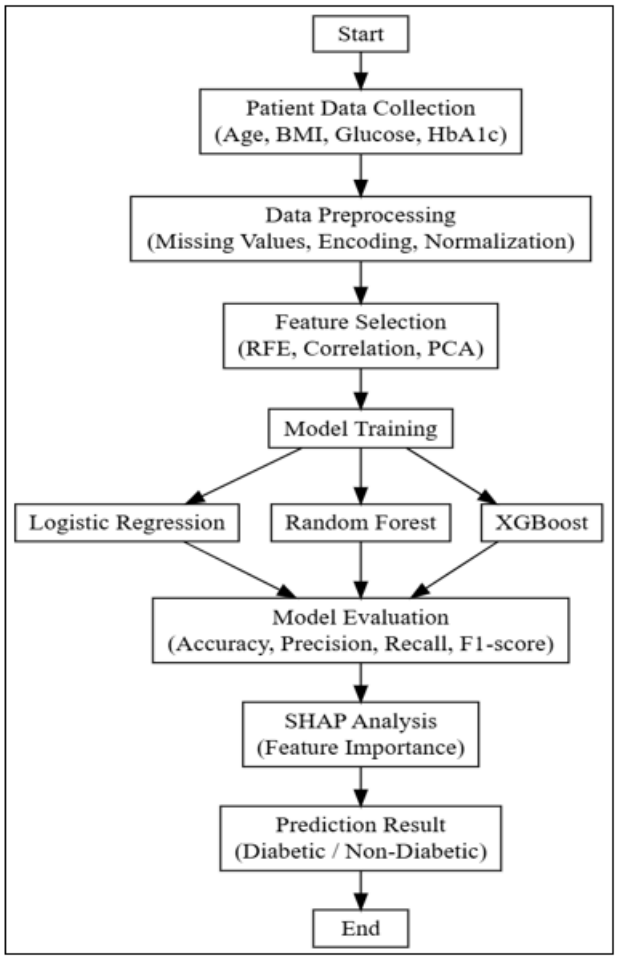

Figure 1: For the proposed system architecture for early diabetes prediction using machine learning.

8. Research Methodology

The experiment adheres to a systematic machine learning paradigm for the early detection of diabetes based on clinical and lifestyle information. The algorithm is realized via Scikit-learn that offers effective and optimized techniques for classification purposes [15]. Besides, XGBoost algorithm is added as a sophisticated boosting algorithm to improve prediction performance.

Preprocessing and transformation of data are done through Pandas and NumPy, which facilitate efficient management and manipulation of structured data. Data visualization is realized using Matplotlib and Seaborn, which provide a clear representation of the patterns in data. Finally, SHAP (SHapley Additive Explanations) is employed to interpret the model outcomes by determining the importance of each feature [12].

9. Dataset Information

The dataset used in this research is obtained from the diabetes dataset available under open-source. This dataset consists of structured information about the health and lifestyle of the patients [19]. Some of the key variables that are present within the dataset include age, gender, BMI, high blood pressure, heart disease, HbA1c, and glucose levels. These variables are used to show whether or not the patient has diabetes [19]. In the light of the presence of several factors that affect the health of the patients within the dataset, this makes it a dependable dataset to be used. In this regard, it can be utilized to create machine learning models that can predict diabetes.

10. Tools and Software

The methodology has been carried out using several reliable software tools and packages. To implement this research process, the Python programming language was chosen since it can be used to carry out many other functions of data science [16]. Libraries such as scikit-learn, xgboost, Pandas, Numpy, Matplotlib, and Seaborn have been used within this research [16]. Interpretability has been achieved through SHAP library [15].

11. Algorithms/Models

In this study, diabetes prediction will be done through application of three types of machine learning algorithms. One important purpose of using several algorithms is to compare their performances and establish which one is best suited for the task.The first algorithm used is Logistic Regression that works perfectly well with binary classification problems [11]. This is because it is very easy to implement and also straightforward to interpret. Moreover, this algorithm is computationally inexpensive. The second algorithm to be considered is Random Forest algorithm that applies the use of multiple decision trees [10]. Unlike the previous model, in this algorithm, prediction is done by aggregating the results of several decision trees.

Finally, XGBoost algorithm, an advanced ensemble learning technique, will be used to perform diabetes prediction. It increases the quality of predictions made by iteratively improving previous mistakes. It has been established to work efficiently for predictive modeling tasks.

Performance of these three models is compared after fitting them. Comparison of their performances will help to select the best performing algorithm among them.

12. Experimental Set-Up

Before the building of models, the dataset goes through some data pre-processing process to make sure that the dataset is of good quality [14]. Where there are missing values, measures are taken to deal with them, while categorical variables are encoded into numeric ones for easy computation [16]. On the other hand, numeric features are normalized for consistency.

Afterward, the dataset is split into two; the first one forms the training dataset while the second one serves as the test data set. Here, an 80:20 percentage ratio is adopted where the former is used in the construction of the models while the latter is used to check their performance.

Lastly, all the selected machine learning algorithms are run under similar experimental conditions. They include logistic regression, random forest, and XGboost among others.

Evaluation Metrics

The effectiveness of the machine learning algorithms is measured by standard classification evaluation metrics. The accuracy metric evaluates the general prediction accuracy. The precision metric measures the proportion of true positive predictions among all positive predictions made by the algorithm. The recall metric measures the rate at which the algorithm correctly identifies true positives. The F1 score metric is a balance of the precision and recall metrics. Furthermore, the ROC-AUC metric evaluates the discrimination power of the algorithm in classifying diabetic patients.

13. Results and Analysis

The proposed model framework is tested using various machine learning classifiers and analyzed based on several performance metrics and visualizations. Results from such analysis prove the effectiveness of each of the models to predict early-onset diabetes based on clinical features and lifestyle factors.The three types of classification techniques utilized in this study include logistic regression, random forest, and XGboost. The performance of the three methods will be assessed using accuracy, precision, recall, F1 score, and ROC-AUC. Out of the three classifiers, XGboost is the most accurate with an accuracy of 92%. In contrast, logistic regression and ra [2]ndom forest have accuracies of 86% and 89% [11]. These results confirm the higher efficiency of ensemble-based classifiers, especially XGBoost, in analyzing and recognizing complex patterns [10]. In order to evaluate the classifier's performance in detail, confusion matrices were used as one of the most common approaches to estimate the number of TP, TN, FP, and FN cases. ROC

curves can be utilized to estimate how well a model recognizes positive class, while the most accurate results are achieved when using XGBoost.

Moreover, a feature importance study is carried out through SHAP (SHapley Additive Explanations), whereby information about the contribution of individual features toward the prediction is provided. According to the study, the main contributing features include the blood glucose level, HbA1c level, BMI, and age.

From all the above-discussed analyses, it can be seen that the suggested feature optimization machine learning approach is characterized by high accuracy while ensuring model interpretability at the same time.

Table 1: Accuracy Evaluation of Gradient Boosting, LightGBM, and CatBoost Models

S. No.	Model	Accuracy (%)
1	Logistic Regression	86%
2	Random Forest	89%
3	XGBoost	92%

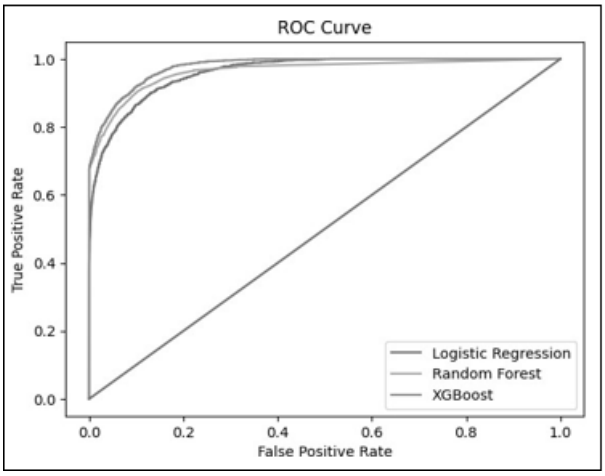

Figure 2: ROC Curve Comparison of Logistic Regression, Random Forest, and XGBoost Models

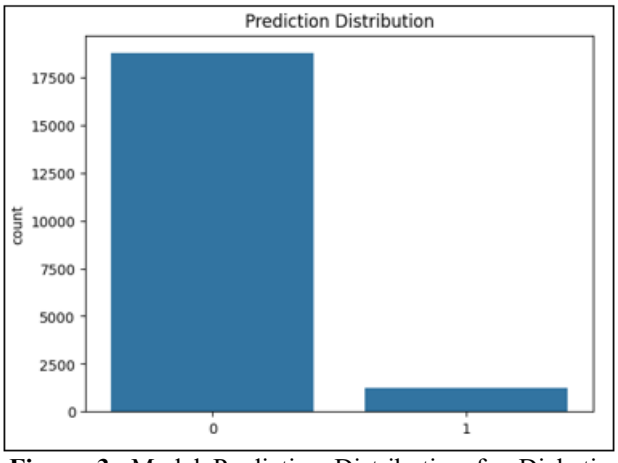

Figure 3: Model Prediction Distribution for Diabetic and Non-Diabetic Classes

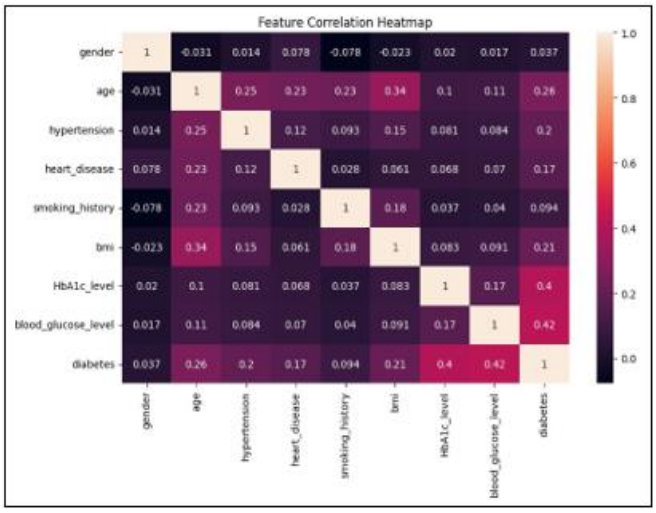

Figure 4: Correlation Heatmap of Features for Diabetes Prediction

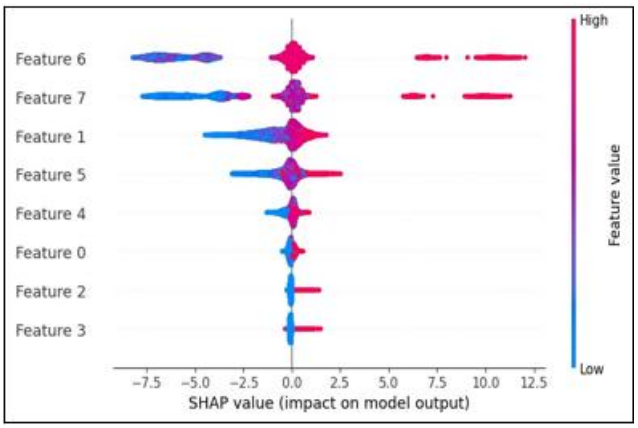

Figure 5: SHAP Summary Plot Showing Feature Importance and Impact on Model Output

14. Discussion

Overall, the performance of the suggested framework is because of choosing features that make sense and have a direct impact on diabetes. They are such indicators as the level of blood glucose, hemoglobin levels, body mass index (BMI), and age, which contribute heavily to the development of diabetes.

In order to achieve better performance, effective data preprocessing was applied in the suggested framework, such as categorical data encoding and numerical data normalization. Thus, the model was able to learn and perform predictions more effectively and stably.

In the experiments, the results demonstrated that ensemble-based models perform better than simple ones owing to their ability to find complex patterns in the data set. As for XGBoost, it showed the highest performance, having reached an accuracy of 92%. The reason lies in its iterative nature when each step eliminates errors and thus improves the quality of predictions.

In the evaluation, different criteria were applied to ensure that the models performed well in various conditions. Apart from accuracy, precision, recall, F1-score, and ROC-AUC are evaluated. Moreover, the SHAP technique was used to

improve model interpretability through finding out individual feature contributions.

Benefits

- Offers high prediction accuracy, where XGBoost and Random Forest yield superior results when compared to other algorithms.
- Employs several machine learning algorithms in analyzing and comparing various models for the purpose of choosing the most accurate algorithm for predicting diabetes.
- Implements methods of feature selection which ensure high-quality data and consequently increase the accuracy of the model through the use of significant variables.
- Aids in determining significant health factors such as blood sugar levels, body mass index, and age, which affect diabetes prediction.
- Assists in early prediction of diabetes risks for purposes of seeking early medical intervention.
- The entire process can be regarded as relatively simple and efficient.

15. Limitations

However, despite all the positive results, there are some limitations to the analysis. Firstly, the dataset used by researchers might not fully account for diversity within the population and thus impact generalization of the model to different healthcare contexts. Secondly, there are some variables that can be important but are not available in the given dataset, such as lifestyle factors, diet, genetics, etc.The models are tested in an experimental setup without being confirmed in practice.Despite providing enhanced interpretation of the model, some people may experience difficulties understanding the SHAP tool.Last but not least, despite their relatively high efficiency, it is worth investigating other models to improve the performance of the algorithms chosen in this paper (Logistic Regression, Random Forest, XGBoost).

16. Conclusion

The research problem associated with the development of early-stage diabetes detection is addressed through building a model based on machine learning, which takes into account the attributes related to lifestyle and health condition. The purpose of the work can be expressed as providing a highly precise and comprehensible solution aimed at predicting diabetes.For this reason, several supervised learning algorithms were employed, such as Logistic Regression, Random Forest, and XGBoost. First, the data was preprocessed to eliminate irrelevant information. Then, the algorithms were analyzed using common metrics, while the explainability of predictions was assessed via the SHAP value.As follows from the experimental analysis, the model with XGBoost showed the best performance in terms of precision, achieving an accuracy of 92%, which is significantly higher than Logistic Regression (86%) and Random Forest (89%). Also, it is important to emphasize the role of such factors as blood glucose, HbA1c level, BMI, and age.This research makes a significant contribution by integrating highly accurate machine learning algorithms with the methods of explainable AI. Thus, an efficient and easily comprehensible framework can be developed for predicting diabetes at an early stage. This method possesses high potential for being utilized in healthcare data-driven solutions and providing medical experts with useful insights.

17. Future Work

In terms of future research, including a larger number of patients with varying demographics would contribute positively towards strengthening the effectiveness and flexibility of the model [7]. Other possibilities could include experimenting with other sophisticated techniques like hybrid approaches and deep learning in order to detect any correlations between the features [8]. The proposed solution can further be extended towards real-time applications, which will ensure dynamic prediction based on continuous health monitoring. Additionally, inclusion of smart healthcare devices would allow for the process of collecting data automatically while providing immediate risk assessment [20].Other aspects that can be included in the prediction process include genetic data, diet, and detailed life style patterns to further boost the prediction accuracy of the model. Simplification of the technique of explaining the results is another area that can be addressed in the future.

References

[1] Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
[2] S. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.
[3] M. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
[4] G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients," International Journal of Engineering and Advanced Technology, vol. 1, no. 3, pp. 147–151, 2012.
[5] H. Wu et al., "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.
[6] J. Perveen et al., "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2016.
[7] M. Chen et al., "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.
[8] R. Miotto et al., "Deep learning for healthcare: Review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[11] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. New York, NY,

USA: Wiley, 2013.
[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
[13] Molnar, Interpretable Machine Learning. Munich, Germany: Lulu Press, 2022.
[14] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.
[15] Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.
[16] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.
[17] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Boston, MA, USA: Pearson, 2018.
[18] World Health Organization, "Global report on diabetes," WHO Press, Geneva, Switzerland, 2016.
[19] Kaggle, "Diabetes Prediction Dataset," Available: https://www.kaggle.com/datasets
[20] E. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.