International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume-10, Issue-6, December 2023
https://doi.org/10.55524/ijirem.2023.10.6.14
Article ID IJIRD-1311, Pages 92-99
www.ijirem.org

Machine Learning Based Diabetes Prediction System: A Novel Approach

Shalini Shekhar[1], and Dr. Nikita Thakur[2]

[1]Research Scholar, Department of Computer Science, Sai Nath University, Ranchi, Jharkhand, India
[2]Associate Professor, Department of Computer Science, Sai Nath University, Ranchi, Jharkhand, India

ABSTRACT- The healthcare sector is poised to experience a remarkable transformation with the integration of artificial intelligence. In the realm of healthcare analysis and prediction, the utilization of data science and machine learning applications proves advantageous. Healthcare is emerging as a progressive and promising field for the implementation of data science applications, particularly in Medical Images Analysis, Drug Discovery, Genetics Research, and Predictive Medicine. Diabetes is broadly classified into three main types: type 1, type 2, and gestational diabetes. The primary objective of this research is to develop a Machine Learning Model for the diagnosis of diabetes. Identifying the accurate symptoms in users or individuals with diabetes is a significant challenge for application and the execution of rules. These combinations of knowledge determine whether an individual is a diabetes patient, including its subtypes such as type_1, type_2, and gestational diabetes. The Machine Learning Model underwent testing on a cohort of 150 patients, producing results comparable to those of medical professionals.

KEYWORDS- Diabetes Prediction System, Machine Learning, Classification Models.

I. INTRODUCTION

The Machine Learning Model crafted by the researcher can be utilized effectively and efficiently for diagnosing various types of diabetes. This is particularly beneficial in less developed countries where the availability of doctors is limited. The intelligent Machine Learning Model aims to decrease reliance on medical professionals, assisting both healthcare providers and patients in making more precise and expedited decisions. Type 1 diabetes results from the destruction of the islets of Langerhans in the pancreas due to autoimmune systems, which are responsible for producing insulin. This type accounts for only five to ten percent of the total diabetic population. The majority of diabetic individuals fall into the Type 2 category, strongly linked with obesity. Type 2 diabetes is characterized by a combination of insulin resistance and insufficient insulin production, constituting ninety to ninety-five percent of the total diabetic population. On the other hand, gestational diabetes is a short-term disorder characterized by elevated blood sugar levels during pregnancy, which typically return to normal after delivery. Women who have experienced gestational diabetes are at a higher risk of developing diabetes in the future.

II. METHODOLOGY

The proposed framework is segmented into various phases. The visual representation of the process is depicted in Figure 1. Python Jupyter Note was utilized for the entire implementation. Different libraries including NumPy, pandas, scikit, and Matplotlib have been employed in the analysis of the data. The tasks executed in each phase and the pertinent functions extracted from Python toolkits are expounded below.

A. Dataset (PIDD)

The Pima Indian Diabetes Database is a well-known and widely used dataset for predicting diabetes. This dataset comprises 768 rows and 9 columns, encompassing attributes such as glucose, pregnancies, skin thickness, blood pressure, BMI, insulin, age, and outcomes. The outcome variable predicts whether the patient is diabetic [19], positive, or diabetic-negative. The Pandas function is utilized to read the CSV file where the dataset file is in Excel format.

B. Data Visualization

Data visualization aids in comprehending the data more effectively by presenting it visually. In this stage, data are represented in the form of a bar chart. The analysis unveils the percentage of people affected by diabetes diseases. It also showcases information from the dataset, including age, blood pressure, pregnancies, and glucose. Additionally, it predicts how many people are affected by diabetes out of 768. For displaying output, graphical representation functions such as plot axis, pyplot, and others are employed.

C. Preprocessing

This section involves the removal of outliers and standardizing the data. The processed data are used for creating a model. The data should be preprocessed and organized properly before applying classifiers to the data index. In this phase, inconsistent data are handled and removed to obtain more precise and accurate results. The dataset contains missing values, and selected attributes like blood pressure, skin thickness, glucose level, and BMI are assigned with missing values, as these parameters cannot

have null values. Subsequently, all values are normalized by scaling the dataset.

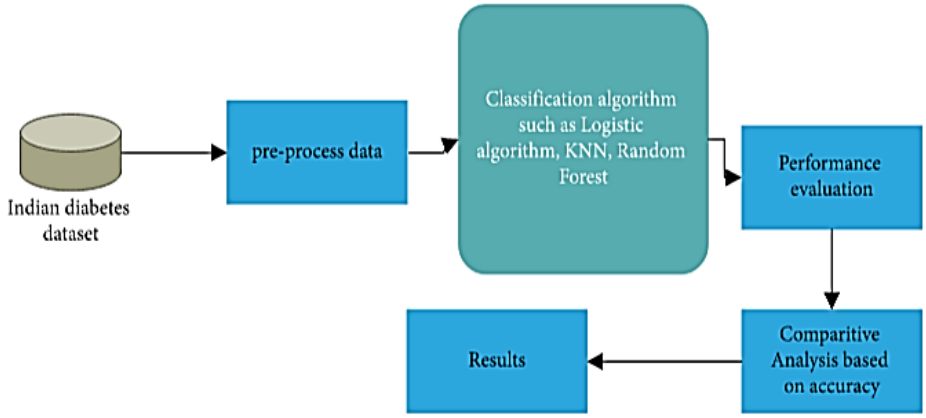

Figure 1: Framework of ML techniques.

D. Machine Learning Classification Algorithms

Following the preprocessing of the data, ML classifiers are employed using the scikit-learn Python Toolkit. Scikit is a straightforward toolkit utilized for processing and analyzing data [20]. These toolkits find application in a majority of tasks. Initially, a function like model selection train-test split is used to divide the dataset into training and testing datasets. Due to the limited dataset source, approximately 90% of the dataset is allocated for training purposes, and the remaining 10% is utilized for testing by randomly selecting the data. Subsequently, various classifiers, such as ML algorithms [21], are implemented for diagnosing diabetes. ML classifiers are chosen due to their simplicity and widespread use. As this study emphasizes hyper parameter tuning, it will be detailed in the subsequent section.

E. Hyper parameter Tuning

Hyper parameter tuning is employed to assess the ML models. The process of selecting an optimal set of hyperparameters is termed hyperparameter tuning [22]. The value of the hyperparameter's model is fixed before commencing the ML task. Hyperparameter tuning plays a pivotal role in ML techniques, where model parameters are derived from the data. To achieve the best fit, hyperparameter tuning is conducted. Selecting the best hyperparameter is a complex problem, hence grid search and random search algorithms are employed. This technique is adopted to enhance the accuracy of the ML classifier [23].

III. MACHINE LEARNING CLASSIFICATION MODELS

A. Logistic Regression (LR)

LR models have been obtained from the field of statistics. This algorithm is utilized for binary classification problem statements. The primary objective of LR is to ascertain the values of coefficients. LR transforms the values into the range of 0-1. The LR model determines the probability of the given data instance belonging to a class and predicts it as either 0 or 1. This approach can be employed for scenarios where there are multiple factors contributing to the prediction.

The standard function of LR is defined as follows:

$$h\theta\,(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}. \quad (1)$$

Equation (1) delineates the logistic decision for the predicted data. Here, X signifies the data label, and the constants are denoted by β1 and β0.

K-Nearest Neighbor (KNN)

KNN stands as one of the supervised learning techniques in machine learning [25]. It finds extensive application in classification problems. KNN functions by classifying objects based on their proximity or distance, specifically the distance between the object and all objects in the training data. The item is classified based on the K-nearest neighbors, where K, a positive integer, is defined prior to executing the algorithm. Frequently, Euclidean distance is employed to compute the various measures of objects [26]. The Euclidean distance calculation equation is provided below:

$$\text{Euclidean} \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}, \quad (2)$$

$$\text{Manhattan} \sum_{i=1}^{k} |x_i - y_i|. \quad (3)$$

Derived from equations (2) and (3), the Euclidean and Manhattan measures of the KNN classifier are computed using the x and y data up to the i variables.

B. Support Vector Machine (SVM)

The SVM algorithm is a supervised machine learning technique [24]. This model is particularly effective for small datasets with minimal outliers. The primary goal is to determine the hyper plane that effectively separates the data points. Once identified, this hyper plane divides the space into distinct domains, each containing similar types of data.

$$\| \mathbf{x} \| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}. \quad (4)$$

Equation (4) describes the decision state of the support vector machine, where a hyper plane divides the space into two sectors. This hyperplane, serving as a binary classifier, is specifically applied to linear classification [27]. The subspace of a single dimension is constrained by its circumstances. Figure 2 illustrates the classification of SVM hyperplanes.

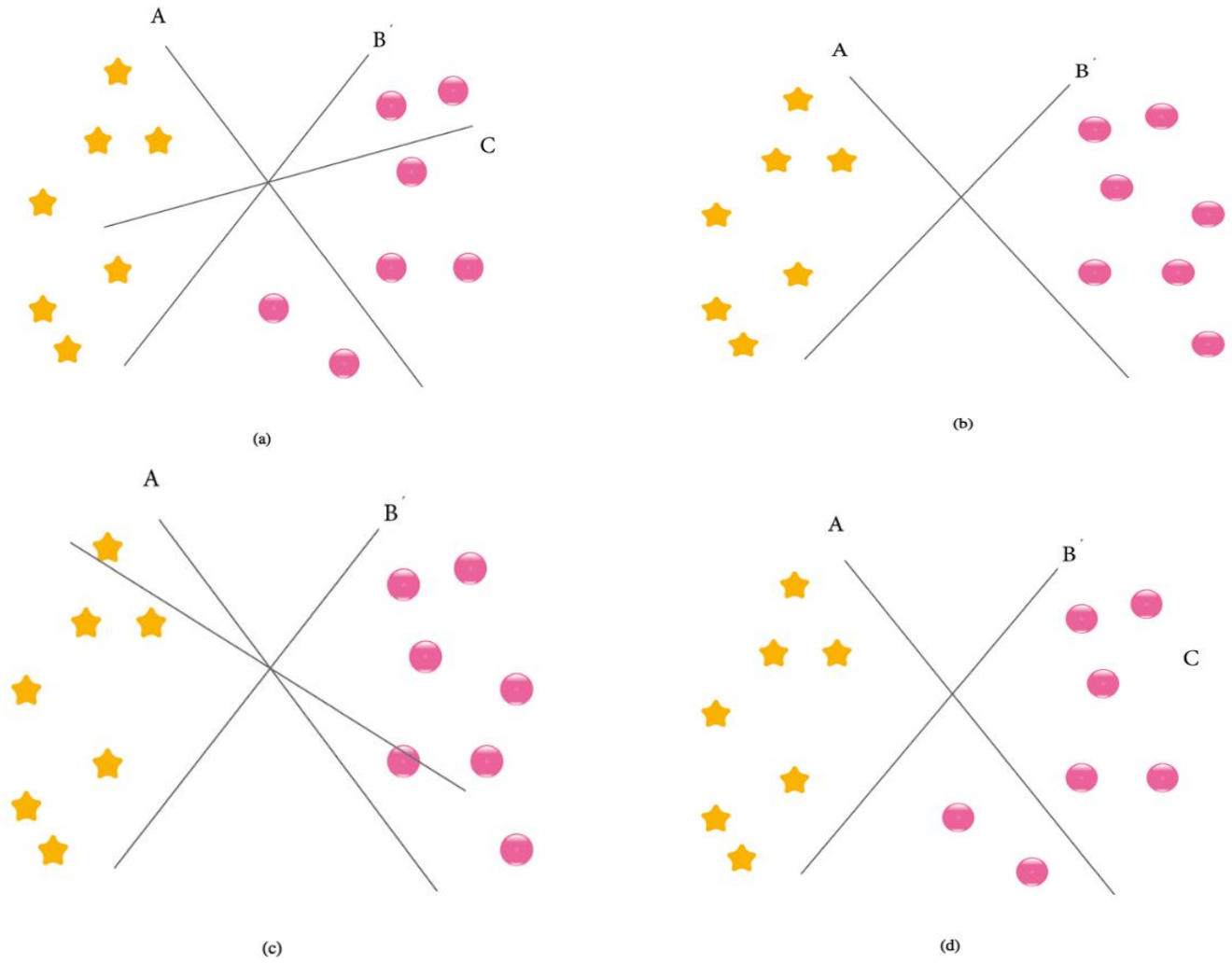

Figure 2: Illustrates the classification of SVM hyperplanes

C. Random Forest

Random Forest is a machine learning algorithm falling under the supervised learning model. The Random Forest classifier is composed of multiple decision trees, each focusing on different aspects of the given dataset. To enhance predictive accuracy [28], it averages the subset results from each tree. Rather than relying on a single decision tree, Random Forest takes a majority vote prediction from all the trees to produce the final output.
Every decision tree node poses a question about the data, contributing to the overall classification process.

IV. PROPOSED FRAMEWORK FOR LOGISTIC REGRESSION

The introduction of hyperparameter tuning is a pivotal component of the proposed framework for Logistic Regression (LR). In the context of machine learning (ML) models, hyperparameter tuning involves parameterizing models, and their behavior can be adjusted based on the problem statement. ML models typically possess various parameters or attributes, and finding the optimal combination of these attributes can be treated as a search problem. For LR tuning, two specific strategies were employed: grid search and random search.

$$\min_{w,b} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \Pr\left(y_i = 1 | \mathbf{x}_i; \mathbf{w}, b\right) + \mathscr{R}\left(\mathbf{w}\right) \right\}$$
$$\mathrm{R}(w) = \lambda \left\{ \alpha \sum_{i=1}^{p} |w_i| + \frac{(1-\alpha)}{2} |\mathbf{w}|^T \mathbf{L} |\mathbf{w}| \right\}. \quad (5)$$

Equation (5) outlines the grid search algorithm utilized in this study. In the context of grid search, the ML model R incorporates hyperparameters x1, x2, and x3. To employ grid search effectively, the values for these hyperparameters (x1, x2, x3) must be defined. The grid

technique involves generating multiple versions of the model R, each with a unique combination of hyperparameter values (x1, x2, x3) pre-determined at the outset. This process allows for the tuning of hyperparameter values in a systematic grid pattern.

Throughout the grid search, individual parameters are isolated, and the optimal potential value is sought while maintaining the other parameters constant. Although this method may result in a less efficient model score compared to random search, it demonstrates enhanced exploratory power. The greater exploratory power is attributed to its ability to systematically search within critical parameter ranges, facilitating the identification of optimal values (hyperparameters). In the scope of this research, grid search is applied to enhance the effectiveness and efficiency [29] of the LR classifier, ultimately improving the accuracy of the prediction model.

A. Results and Evaluation

The Pima Indian Diabetes Database (PIDD) dataset comprises 768 entries, with 500 patients identified as nondiabetic. A visual representation of the comparison of the proposed technique is presented in Figure 3 using a bar chart.

Following the completion of data processing, the training dataset is partitioned, and four machine learning (ML) classifier algorithms are deployed. To achieve optimal results for the given dataset, hyperparameter tuning and cross-validation are executed. As previously detailed, K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) are the ML algorithms applied in this study. The subsequent sections delve into the specifics of hyperparameter tuning, and the outcomes obtained from all the models are elaborated upon.

The performance of the ML algorithms is evaluated using various metrics, including B1 score, recall, precision, and accuracy [30]. The equation governing these evaluation metrics is provided below:

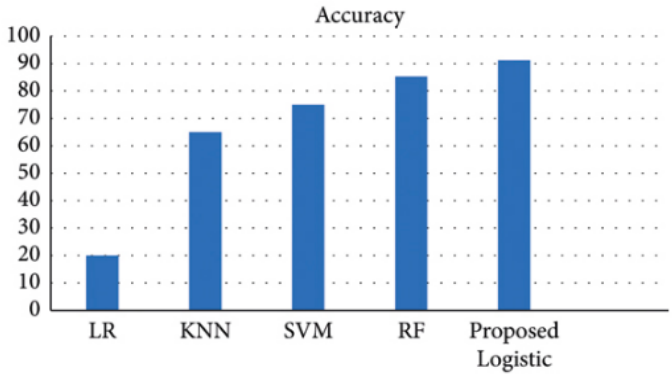

Figure 3: Representation of the bar chart of this data set.

Following the data processing phase, the training dataset is partitioned, and four machine learning (ML) classifier algorithms were employed. To attain optimal results for the provided dataset, hyperparameter tuning and cross-validation procedures were executed. As elaborated earlier for ML algorithms, the techniques applied encompassed K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). The subsequent sections delve into the specifics of hyperparameter tuning, presenting the results obtained from all models. The performance of the ML algorithms is thoroughly assessed using diverse evaluation metrics, including B1 score, recall, precision, and accuracy [30]. The corresponding equation is provided below:

$$\begin{aligned} \text{sensitivity} &= \frac{TP}{TP+FN} \times 100, \\ \text{specificity} &= \frac{TN}{TN+FP} \times 100, \\ \text{accuracy} &= \frac{TP+TN}{TP+FP+TN+FN}, \\ \text{PPV} &= \frac{TP}{TP+FP}. \end{aligned} \tag{6}$$

(i) ***True negative (TN):*** This occurs when the actual classification is False (F), and the predicted samples are also correctly classified as False (F).

(ii) False positive (FP): In this scenario, the actual classification is False (F), but the predicted samples are incorrectly classified as True (T).

(iii) False negative (FN): This happens when the actual classification is True (T), but the predicted samples are incorrectly classified as False (F).

B. Analysis of Results Using Various ML Techniques

This study involved the construction of four classifier models, namely LR, RF, SVM, and KNN. As part of the data preprocessing, outliers were removed before training the *models*. The comparison of ML algorithms is visually represented in a bar chart. The analysis, depicted in Figure 3, highlights that RF and SVM achieved a commendable accuracy of 83%. The application of hyperparameters to LR resulted in a notable 3% improvement in the accuracy of predictions [31].

Figure 4 illustrates the correlation of the confusion matrix, which simultaneously addresses missing values and outlier rejection. The correlation attribute with the target variable reveals a substantial improvement in the correlation coefficient, summarized through statistical data presented in a box plot. This statistical summary includes key metrics such as maximum, minimum, first quartile, median, and third quartile.

Addressing null values in BMI and blood sugar is crucial during the data preprocessing stage. An examination of BMI and pregnancies, as shown in Figure 5, reveals a robust positive correlation between BMI and the number of pregnancies [32]. Diagnosed diabetic-positive individuals tend to have a higher BMI compared to nondiabetic individuals, with minimal differences among medians. Generally, women with a higher number of pregnancies exhibit elevated BMI [33].

Furthermore, the relationship between pedigree function and clinical test reports indicates that individuals with a high pedigree function are more likely to test positive, while those with low pedigree function tend to test negative.

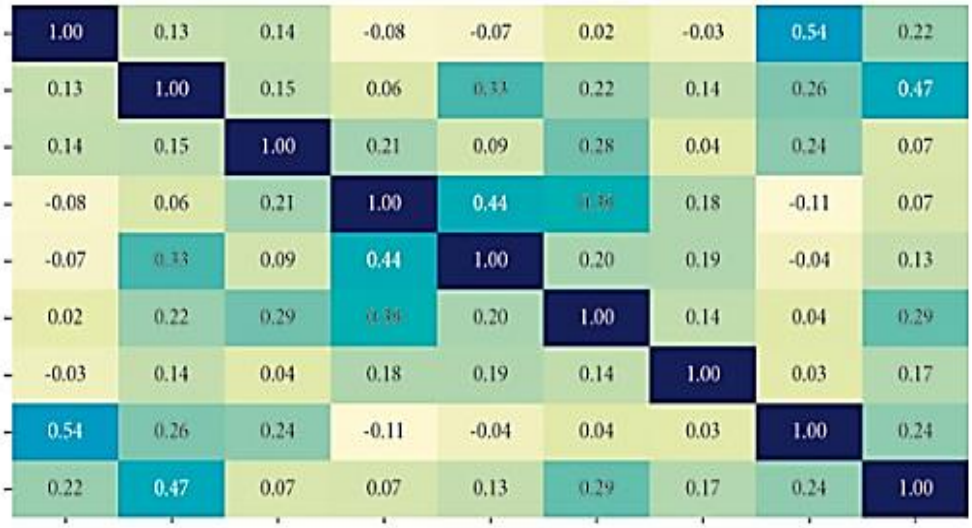

Figure 4 :Analysis of correlation between variables.

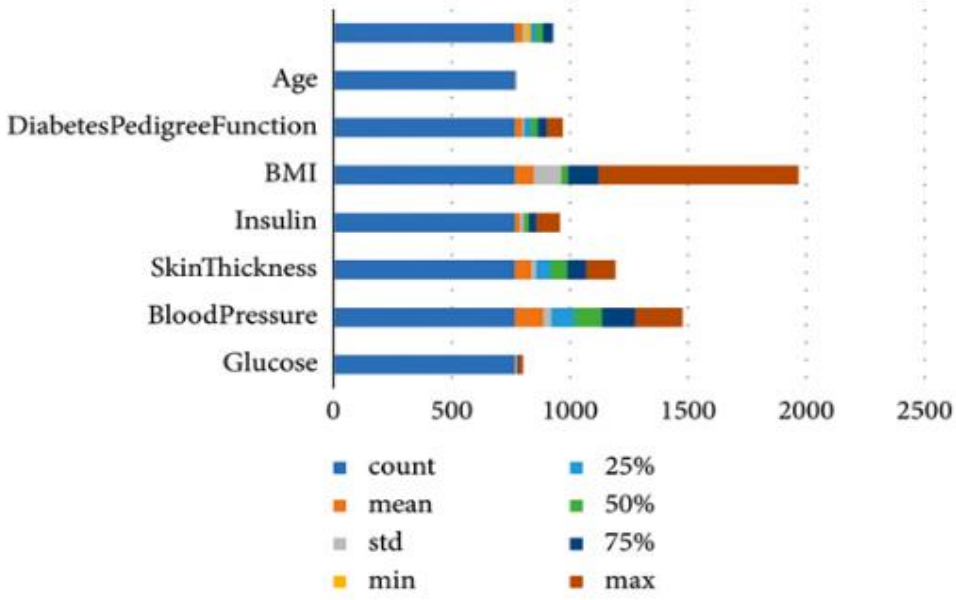

Figure 5: Positive correlation between BMI and the number of pregnancies

Given that individuals who tested positive exhibit a higher median and outliers, the pedigree function proves instrumental in accurately estimating diabetic test results. This observation underscores the hereditary nature of diabetes, suggesting a substantial genetic component in the development of diabetes within the PIMA Indians Diabetes dataset. Figure 6 visually demonstrates a notable disparity in the average number of pregnancies, with diabetic women having a higher average (4.9) compared to nondiabetic women (3.3).

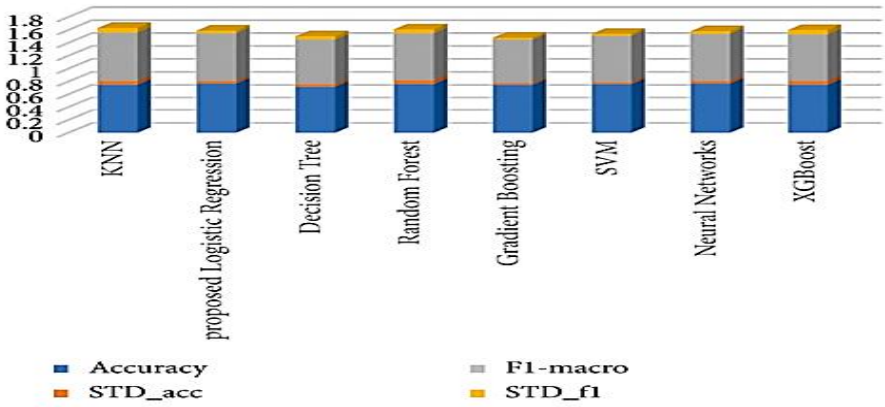

Figure 6: Existing association between the test result and the pedigree function.

Figure 7 illustrates that women with a normal weight face a ninefold increase in the risk of diabetes diagnosis compared to overweight women. The BMI is notably elevated within the interquartile range for women who tested positive for diabetes.

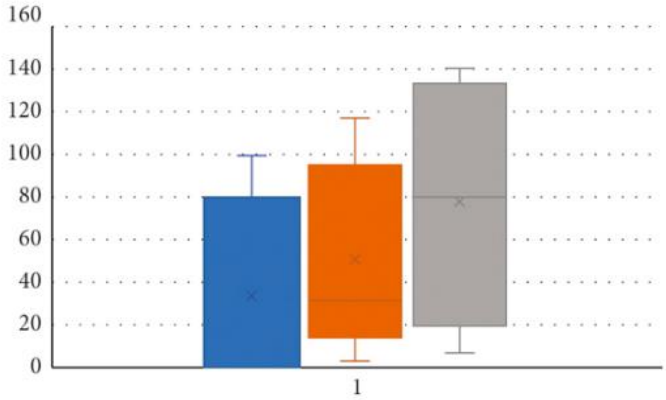

Figure 7: BMI had a close association with the occurrence of diabetes.

Women in the age group above 31 years face a heightened risk of being diagnosed with diabetes compared to those in the younger age group.

Figure 8 presents the confusion matrix for the proposed work. The diagonal elements of the matrix classify the number of data points for each class. Accuracy is calculated by summing the components on the diagonal and dividing it by the sum of elements in the entire matrix. A model is deemed effective when the confusion matrix has larger values on the main diagonal and smaller values elsewhere.

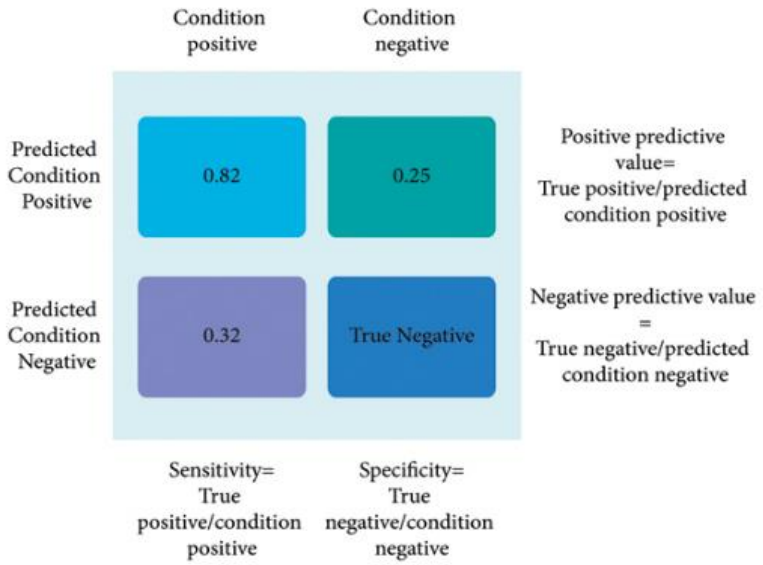

Figure 8: Confusion matrix of ML algorithms

The receiver operating characteristic (ROC) plot serves as a valuable tool for assessing algorithm performance, with successful applications in healthcare prognosis and diagnosis. A system or model is deemed effective when the reference point emphasizes the upper left corner of the ROC chart [34]. This reference point signifies high sensitivity and fewer false-positive values. The area under the ROC curve (AUC) is a reliable normalization metric, where an AUC above 0.5 indicates a robust test method. Figure 9 depicts the ROC value of LR, achieving a notable 86%, surpassing others [35]. This analysis suggests that RF is well-suited for accurately predicting diseases.

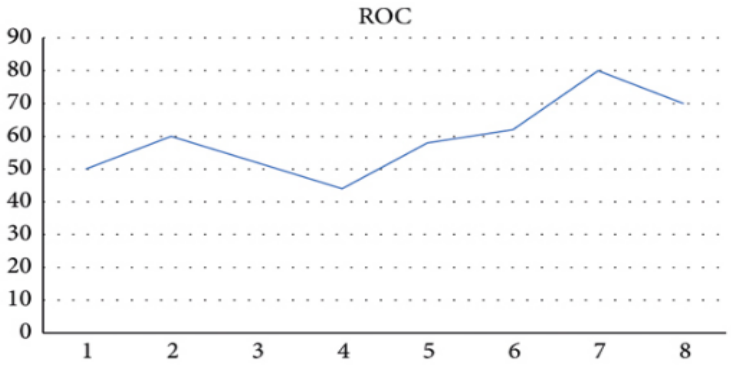

Figure 9: ROC of logistic regression.

Figure 9 illustrates the ROC curve for the proposed logistic regression. The ROC performance improves during the training phase, although some discrepancies in the training data may lead to errors.

V. CONCLUSION

In summary, machine learning (ML) techniques prove valuable in disease diagnosis, particularly for early detection that facilitates prompt medical attention. This study delves into existing ML classification models for predicting diabetic patients, prioritizing accuracy as a key metric. The ML technique is applied to the Pima Indian Diabetes Database (PIDD) dataset, trained, validated on the test dataset, and subsequently verified.

The findings from our implementation highlight the superior performance of Logistic Regression (LR) compared to other ML algorithms. Notably, glucose and BMI exhibit a robust correlation with diabetes, as indicated by association rule mining. The ROC value for LR is determined to be 86%. It is essential to acknowledge a limitation in the study, namely the reliance on a structured dataset, with future considerations aiming to incorporate unstructured data.

The scope of future research involves extending the models' applicability to other healthcare domains, enabling predictions for diseases such as cancer, Parkinson's disease, heart disease, and COVID-19. Additionally, there is potential to enhance the predictive capabilities of diabetes by considering additional attributes such as family history, smoking habits, drinking habits, and physical inactivity.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Alama, T. M., Iqbala, M. A., Ali, Y., et al. (2019). A Model for Early Prediction of Diabetes. Informatics in Medicine Unlocked, 16, Article ID 100204.

[2] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. In Proceedings of the 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, UK, September 2018.

[3] Mahabub, A. (2019). A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques. SN Applied Sciences, Springer.

[4] Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An improved artificial neural network model for effective diabetes prediction. Complexity, 2021, Article ID 5525271.

[5] Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm. Health Information Science and Systems, 8.

[6] Ahmed, M. H., Elghandour, M. M. Y., Salem, A. Z. M., et al. (2015). Influence of Trichoderma reesei or Saccharomyces cerevisiae on performance, ruminal fermentation, carcass characteristics and blood biochemistry of lambs fed Atriplex nummularia and Acacia saligna mixture. Livestock Science, 180, 90–97.

[7] Daliri, M. R. (2012). Automatic diagnosis of neuro-degenerative diseases using gait dynamics. Measurement, 45(7), 1729–1734.

[8] Dwivedi, K. (2019). Analysis of decision tree for diabetes prediction. International Journal of Engineering and Technical Research, 9.

[9] Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4), 702–710.

[10] Liu, C., Zoph, B., Neumann, M., et al. (2018). Progressive neural architecture search. In European Conference on Computer Vision (ECCV), pp. 19–34, LNCS Springer, Munich, Germany.

[11] Anouncia, M., Lj, C. M., Jeevitha, P., & Nandhini, R. T. (2013). Design of a diabetic diagnosis system using rough

sets. Cybernetics and Information Technologies, 13(3), 124–139.
[12] Valdez, P. J., Tocco, V. J., & Savage, P. E. (2014). A general kinetic model for the hydrothermal liquefaction of microalgae. Bioresource Technology, 163, 123–127.
[13] Muthukaruppan, S., & Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications, 39(14), Article ID 11657.
[14] Ganji, M. F., & Abadeh, M. S. (2011). A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12), Article ID 14650.
[15] Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer Methods and Programs in Biomedicine, 104(3), 443–451.
[16] Zou, Q., Qu, K., Luo, Y., & Yin, D. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9.
[17] Wang, W., Meng, T., & YU, M. (2020). Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization. IEEE Access, 8, 217908–217916.
[18] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8.
[19] Kapoor, S., & Priya, K. (2018). Optimizing hyper parameters for improved diabetes prediction. International Research Journal of Engineering and Technology, 5.
[20] Srivastava, S., Sharma, L., Sharma, V., & Kumar, A. (2020). Prediction of diabetes using artificial neural network approach. In Engineering Vibration, Communication and Information Processing, 29, Springer, Berlin/Heidelberg, Germany.
[21] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47.
[22] Nai-aruna, N., & Moungmaia, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69.
[23] Mujumdara, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165.
[24] Roy, V., Shukla, P. K., Gupta, A. K., Goel, V., Shukla, P. K., & Shukla, S. (2021). Taxonomy on EEG artifacts removal methods, issues, and healthcare applications. Journal of Organizational and End User Computing, 33(1), 19–46.
[25] Khambra, G., & Shukla, P. (2021). Novel machine learning applications on fly ash based concrete: an overview. Materials Today Proceedings.
[26] Shukla, P. K., Sandhu, J. K., Ahirwar, A., Ghai, D., Maheshwary, P., & Shukla, P. K. (2021). Multiobjective genetic algorithm and convolutional neural network based COVID-19 identification in chest X-ray images. Mathematical Problems in Engineering, 2021, Article ID 7804540.
[27] Rathore, N. K., Jain, N. K., Shukla, P. K., Rawat, U. S., & Dubey, R. (2021). Image forgery detection using singular value decomposition with some attacks. National Academy Science Letters, 44, 331–338.
[28] Agrawal, M., Khan, A. U., & Shukla, P. K. (2019). Stock price prediction using technical indicators: a predictive model using optimal deep learning. International Journal of Recent Technology and Engineering (IJRTE), 8(2), 2297–2305.
[29] Roy, V., Shukla, S., Shukla, P. K., & Rawat, P. (2017). Gaussian elimination-based novel canonical correlation analysis method for EEG motion artifact removal. Journal of Healthcare Engineering, 2017, Article ID 9674712.
[30] Gupta, R., & Shukla, P. K. (2015). Performance analysis of anti-phishing tools and study of classification data mining algorithms for a novel anti-phishing system. International Journal of Computer Network and Information Security (IJCNIS), 7(12), 70–77.
[31] Kumar Ahirwar, M., Shukla, P. K., & Singhai, R. (2021). Cbo I E.: A Data Mining Approach for Healthcare IoT Dataset Using Chaotic Biogeography-Based Optimization and Information Entropy. Scientific Programming, 2021, Article ID 8715668.
[32] Bhatt, R., Maheshwary, P., Shukla, P., Shukla, P., Shrivastava, M., & Changlani, S. (2020). Implementation of fruit fly optimization algorithm (FFOA) to escalate the attacking efficiency of node capture attack in wireless sensor networks (WSN). Computer Communications, 149, 134–145.
[33] Ojesina, A. I., Lichtenstein, L., Freeman, S. S., et al. (2014). Landscape of genomic alterations in cervical carcinomas. Nature, 506(7488), 371–375.
[34] National Heart Lung Blood Institute. (1995). In National Institute of Diabetes, Digestive, & Kidney Diseases (Us), National Heart, Lung, Blood Institute, Bethesda, MA, USA.
[35] Mather, H. M., Nisbet, J. A., Burton, G. H., et al. (1979). Hypomagnesaemia in diabetes. Clinica Chimica Acta, 95(2), 235–242.