International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Optimizing Diabetes Prediction: LDA Pre-processing & ANN Classification in Healthcare`

Soumya K N[1], Raja Praveen K N[2]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: Diabetes mellitus (DM) is a chronic disease that poses significant health risks if not well managed. The current healthcare system is overwhelmed by the impact of DM. Modern machine learning and deep learning methods have a hard time correctly predicting the stages of diabetes and often encounter decreased classification accuracy when dealing with massive datasets. In this work, we provide a novel approach to address these problems by integrating pre-processing with LDA and ANN for classification. By combining the LDA and ANN probability distribution functions by back propagation with initialized weights, our method enhances the accuracy of diabetes categorization. After pre-processing data from the PIMA and NCSU datasets using min-max normalization, bivariate filter-based feature selection is used to identify crucial characteristics. Pearson correlation is used to improve the feature set according to a threshold value, further refining the selected qualities. Our experimental results demonstrate the efficacy of the proposed approach, surpassing even the most cutting-edge methods. By integrating a robust classification model with advanced pre-processing techniques, our strategy produces encouraging outcomes in the accurate prediction of diabetes, which in turn helps to improve healthcare management methods.

Keywords: *Diabetes Mellitus (DM), Latent Dirichlet Allocation (LDA), Normalization, Pre-processing techniques, Classification accuracy.*

1. Introduction

Chronically high blood glucose levels are a defining feature of Diabetes Mellitus (DM), a serious worldwide health issue [1]. It has significant dangers, such as increased risk of cardiovascular disease, stroke, and ` kidney failure [2,3]. The global prevalence of diabetes mellitus (DM) is estimated to be 382 million people as of right now; by 2035, that figure is expected to have doubled to a startling 592 million. This increase emphasizes how urgent it is to implement efficient illness management plans and how important early prediction is to reducing the negative consequences on public health. As a result of this growing epidemic, scientists are using more advanced computational methods specifically, deep learning and machine learning (ML) to analyze massive datasets and create diabetes prediction models [4]. These methods show potential for assessing several characteristics linked to the illness, enabling precise forecasting and categorization. In this context, the Pima Indian Diabetes Dataset's vast collection of information about people with diabetes makes it a priceless tool for developing and testing machine learning models [5]. Additionally, new opportunities for automating intelligent activities in illness detection and management have been made possible by the convergence of machine learning (ML) and artificial intelligence (AI), notably in fields like computer vision. In this study, we aim to forecast the shift from pre-diabetes to diabetes using machine learning (ML). To achieve strong internal and external validation, we will use electronic medical records (EMR) from The Health Improvement Network (THIN) database [6].`

Our goal is to give tailored treatment plans and actionable information about optimum dietary needs for patients at risk of developing diabetes to healthcare practitioners, going beyond simple prediction. Through the integration of clinical knowledge and data-driven methodologies, our aim is to improve patient outcomes and advance the larger objective of improving public health. Nevertheless, there are several difficulties in using ML in the healthcare industry [7, 8]. While individualized risk assessment and early detection are two benefits of machine learning, challenges like data dependability and model interpretability need to be carefully considered [9]. In order to address these issues and optimize the performance of our prediction models, we provide an extensive pre-processing methodology. This method uses standardization, imputation of missing values, and outlier detection to guarantee data quality and improve prediction accuracy.`

To further improve model performance and refine the feature set, our work uses sophisticated feature selection approaches including Pearson correlation and the Bivariate filter method [10]. Through the methodical integration of these methodological improvements, our goal is to surmount the drawbacks of conventional machine learning techniques and create reliable prediction

[1] *Research scholar, School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bengaluru, 562112, India Soumya.kn16@gmail.com*

[2] *Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, 562112, India rajapraveen.k.n@gmail.com`*

models that can precisely anticipate the course of diabetes. In decision, our work is a concentrated attempt to use machine learning (ML) to forecast the course of diabetes, ultimately leading to better patient outcomes and increased public health. This article's next parts will examine related works, methods, findings, and conclusions in more detail, giving readers a thorough understanding of our study project and its implications for the area of diabetes care.`

2. Literature Survey`

Chang, V., Bailey et al. [1] offered a helpful method for classifying and predicting diabetes using machine learning techniques. They employed data cleaning and one-hot encoding as pre-processing techniques, followed by data augmentation and oversampling to equalize class values. However, their approach proved to be unreliable in estimating the chance of the sickness materializing.`

Aiswal, S. et al. [2] proposed a multistage ensemble approach using J48, Random Forest, K-Nearest Neighbor, JRip, and Naïve Bayes classifiers to predict diabetes. They used the ensemble method, but even in complex classification patterns, their technique had accuracy problems. Despite being effective, their technique proved unreliable for large and diverse datasets.`

Yuan et al. [3] created a stacking ensemble method that is optimal for forecasting the course of diabetes. Their technique regarded incorrect prediction findings, crucial for a life-threatening condition like diabetes, even when they used layered classifiers like logistic regression, support vector machines, and k-nearest neighbors. `

Kaul, S. and Kumar, Y. [4]suggested a hybrid ensemble learning strategy to identify diabetes early on by utilizing a cross-validation-based super learner. Even though it was effective, the use of several machine learning algorithms as base learners increased the complexity of the study.`

Assegie, T.A. and Nair, P.S. [5] presented a cuttlefish algorithm-based bioinspired machine learning technique for identifying type 2 diabetes. Despite feature extraction and classification with various classifiers, their approach overlooked computation cost considerations.`

Cahn, A. et al. [6] created an e-diagnostic system for diabetes diagnosis that makes use of ML algorithms. Although successful, their approach highlighted the significance of improving the prognosis of diabetes mellitus and tackling non-communicable illnesses.`

Li, Y.-H. et al. [7] for precise diabetes prediction, the Twice-Growth Deep Neural Network (2GDNN) was presented. Their model demonstrated a slight advantage in managing the severity of diabetes, even after integrating regression approaches and data preprocessing.`

Nibareke, T. and Laassiri, J. [8] suggested a diagnostic approach for diabetes using six ML algorithms and an ensemble classifier. Their method provided global and local explanations for model predictions, enhancing clinical understanding.`

Jaiswal, V., Negi, A. and Pal, T., [9] deployed a number of machine learning strategies, as well as a semi-supervised model that utilized high gradient boosting, for the purpose of diabetes prediction. In spite of the fact that they have addressed class imbalance, their method has to be improved further by including fuzzy logic techniques.`

Annamalai R and Nedunchelian [10] created the OWDANN method, which uses an optimal weighted deep artificial neural network to forecast diabetes and estimate severity levels. Their method successfully corrected noise and repaired damaged data, but further adjustments are needed to make accurate forecasts.`

3. Methodology`

The steps involved in classifying diabetes using the suggested method are thoroughly described in this section, along with the processing sequence. The processes in the classification process include gathering data, pre-processing, feature selection, and diabetic categorization. Initially, pre-processing is done to eliminate characteristics that are improper or unnecessary once the data is taken from one of the publically accessible datasets. Following this phase, the process of feature selection begins in order to identify the relevant and non-redundant characteristics. Finally, the suggested classifier is used to carry out an efficient classification. The process for classifying diabetes is fully shown diagrammatically in figure 1, which is shown below.`

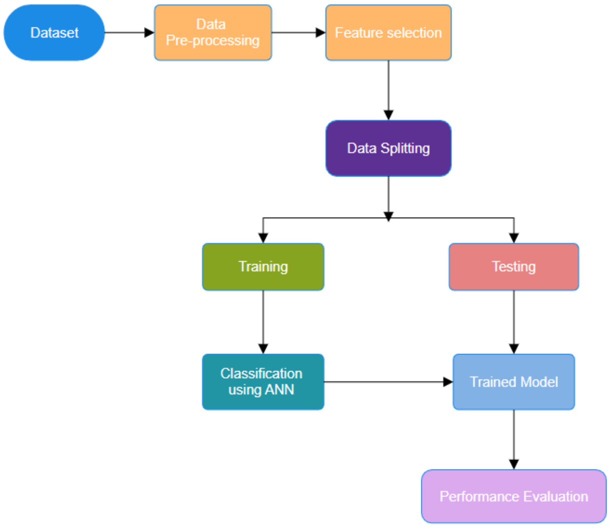


Fig 1. Flow diagram of the proposed method

1. Data Collection`

The PIMA Dataset and the North California State University (NCSU) Dataset were the two separate datasets that supplied the raw data for this investigation. The UCI repository database contains the PIMA Dataset, which is a widely accepted reference point for research on diabetes categorization [6]. The seventy-eight cases comprise nine criteria, including important ones like body mass index, pregnancy, blood pressure, glucose, skin thickness, and insulin (BMI), age group, and the role of the diabetes lineage. Because of its all-inclusiveness and binary outcome variable, it is especially well-suited for supervised learning techniques, such as logistic regression.`

However, the NCSU datase which was acquired from NC State University deepens the investigation because of its distinct qualities. There are 442 examples in all, and the feature set includes things like blood glucose levels, age, sex, and BMI. This dataset offers further insights into diabetes classification and prediction by presenting more elements to look at. Our project ensures a thorough and reliable approach to data collection by utilizing these two databases. Due to the inclusion of both datasets, a thorough study of the many variables impacting diabetes can be conducted, which lays a strong basis for further analysis and classification activities.`

2. Data Pre-processing`

Several crucial actions are taken during the data pre-processing phase of this study in order to get the data ready for further analysis and categorization. To guarantee the dataset's correctness and dependability after the data collecting stage, the first priority is to identify and eliminate outliers. Extreme data points known as outliers are found and removed in order to prevent them from distorting the outcomes of further analysis [4, 5]. To guarantee correctness and completeness, the pre-processing step also involves adding missing values to the dataset. Missing data points may have a significant influence on predictive models' performance, hence it's important to impute or replace them using appropriate techniques. To maintain the dataset's integrity and accuracy, missing value filling is done in this study. Moreover, the variables in the dataset are normalized by the use of standardization. By ensuring that all variables are on a uniform scale, which is usually between 0 and 1, standardization makes it simpler to compare and analyze the data.`

The pre-processing step attempts to lessen the influence of variations in size and magnitude among the characteristics in the dataset by normalizing the variables [8,9]. In general, the pre-processing phase of the data is essential for guaranteeing the quality and consistency of the information, providing the foundation for precise and trustworthy analysis and categorization of diabetes. The data is converted into a format that has been processed and is prepared for further analysis via outlier elimination, missing value imputation, and standardization.`

2.1.1 Outlier identification and removal`

A crucial step in the pre-processing stage of diabetes prediction is the detection and elimination of outliers, which addresses extreme data points that might skew the study and improve model accuracy and dependability. Data points that substantially deviate from the majority are known as outliers, because they have the ability to distort the model's results and provide inaccurate predictions. Thus, in order to guarantee the model's robustness and generalizability, outliers must be found and eliminated.`

Pre-processing methods are used in this study to identify and remove outliers, with a particular emphasis on multivariate analysis to efficiently find outliers across many dimensions. The Mahalanobis distance, as shown in Equation

$$D_M = \left(X - \underline{X}\right) S^{-1} \left(X - \underline{X}\right)`$$

is a frequently used technique that calculates the separation between each observation and the average value of the original variables. Large Mahalanobis distance observations are categorized as outliers and are then removed from the dataset.`

Furthermore, a multivariate method like Principal Component Analysis (PCA) is used in high-dimensional datasets where outliers may not be immediately visible in particular dimensions. Through its analysis of the data's underlying structure and identification of patterns that may point to outliers, PCA is a useful method for successfully finding and removing outlier observations.`

All things considered, the procedure of locating and eliminating outliers is essential to improving the dataset's quality for diabetes prediction. The pre-processed data becomes more dependable and appropriate for further analysis by methodically locating and removing outliers, which eventually results in more accurate and dependable forecasts.`

2.1.2 Filling missing values`

To ensure the dataset is complete and prepared for analysis, filling missing values is a crucial part of the diabetes prediction pre-processing stage. Missing data may have an impact on the precision and dependability of diabetes diagnosis and therapy, which can significantly reduce the efficacy of prediction models. To maintain the prediction model accurate and dependable, missing values must be handled. They could come from a number of factors, such as inadequate or erroneous data entry.`

The common and straightforward approach of imputation missing values in this research project applying the attribute mean values helps to retain the overall structure of the dataset. No critical information is lost and the dataset is maintained intact by replacing the mean for missing values. When dealing with numerical data, this strategy is highly beneficial as it makes it possible to continue the analysis without having to toss away instances that aren't complete.`

Furthermore, the choice to employ mean imputation is compatible with the purpose of producing exact projections for diabetes diagnosis and treatment. The imputed values closely match the properties of the dataset by making advantage of the central tendency of the available data, which results in more accurate projections. The aims of this project are best fulfilled by mean imputation as it strikes a balance between simplicity and effectiveness, even if more complicated approaches like regression imputation or K-Nearest Neighbors (KNN) imputation may give better accuracy.`

This work assures that the dataset is complete and well-prepared for analysis by resolving missing values by mean imputation during pre-processing, giving a solid base for accurate and insightful predictions in diabetes classification, as represented in equation (3). Imputation with the mean offers benefits as it fills in continuous data without introducing outliers.`

$$\boldsymbol{(x)} = \boldsymbol{f(x)} = \{^{\boldsymbol{mean(x),if\ x = null/missed}}_{\boldsymbol{x,otherwise}}$$

as represented in equation (3). Imputation with the mean offers benefits as it fills in continuous data without introducing outliers.`

2.1.3 Standardization`

In the pre-processing phase of diabetes prediction, standardization is essential for normalizing the data and guaranteeing that characteristics are scaled consistently for better prediction model accuracy [4]. A popular method for rescaling attributes to a typical normal distribution with a standard deviation of one and a mean of zero is called Z-score normalization, or standardization. By standardizing the data, we eliminate the impact of various measurement units and enable fair comparisons between them by bringing all attributes to the same scale.`

The standardization procedure is represented by equation (4), $\boldsymbol{R(x)} = \frac{\boldsymbol{x - \underline{x}}}{\boldsymbol{\sigma}}$

However, standardizing features in certain ML models like tree-based models might not always provide significant gains in speed.By ensuring that the standardized values have a zero mean and a one standard deviation, this transformation makes it simpler to compare and understand feature distributions. Additionally, normalization aids in lessening the data distribution's skewness, improving its symmetry and centering it around the mean. This is especially helpful for prediction models since it guarantees that the data complies with the presumptions that underpin several statistical methods, like logistic and linear regression. We can increase the effectiveness of these models and their precision in predicting diabetes outcomes by standardizing the data. A crucial stage in the pre-processing pipeline for diabetes prediction is standardization via Z-score normalization. Standardization increases the overall efficacy and dependability of prediction models by bringing characteristics to a consistent scale and minimizing skewness. This eventually results in more precise and clinically useful predictions for the diagnosis and management of diabetes.`

`2.1.4 Feature selection`

The feature selection process in this project is critical for enhancing the classification accuracy of diabetes prediction models. To properly aid in the classification process, it entails determining which characteristics from the pre-processed dataset are the most relevant. `

In the proposed approach, the feature selection method utilizes a Bivariate statistics approach, specifically employing a Bivariate filter for feature extraction. This filter integrates heterogeneous data layers, addressing uncertainties in the input data. It evaluates the relevance of features by utilizing a certainty factor, which is determined using conditional and prior probabilities. Positive outcomes represent a rise in the certainty value, whilst bad outcomes indicate a fall. In addition, weights are assigned to characteristics depending on the relevance of each one using the Weight of Evidence (WoE) based on Bayesian probability theory. `

Finding factors that substantially influence the categorization of diabetes is the goal of the feature selection method. Then, further analysis is conducted using these chosen attributes as input, such as Pearson correlation, to refine the feature set based on a specified threshold value. This iterative process ensures that only the most relevant and informative features are retained for classification, enhancing the diabetes prediction models in this study, eventually increasing their precision and consistency.`

2.1.5 Pearson Correlation `

The settings are adjusted to minimize redundant data and enhance the association between the Pearson correlation and the features of diabetes. The Pearson correlation coefficient may be used to look at linear correlations between random variables. The linear correlation between two continuous variables is illustrated in the following equation.`

$$r_{xy} = \frac{\Sigma(x_i - \underline{x})\, \Sigma y_i - \underline{y})}{\sqrt{\Sigma(\, x_i - \underline{x})^2}\, \sqrt{\Sigma(\, y_i - \underline{y})^2}}`$$

With an R-value of 0.12, the Pearson correlation shows a less significant relationship with diabetes. It is discovered that there is a modest correlation between a few characteristics and diabetes (r = 0.33, r = -0.42, r = 0.23). It is crucial to remember that a connection does not indicate a cause. With the use of this data, pertinent characteristics that are highly linked to diabetes are found and chosen to serve as input factors for accurate illness categorization. In order to identify type II diabetes patients, the result of Pearson correlation is subsequently included into the classification procedure.`

3. ANN Classification

Following the feature selection phase, the PIMA and NCSU dataset were used for the classification. One of the ML algorithms employed for classification is Artificial Neural Network (ANN), known for providing more accurate results compared to existing techniques. An ANN is made up of one or more hidden layers where information is processed by neurons. In order to get better outcomes, every node serves as an activation node that classes the output of artificial neurons. To minimize the training error, several parameters can be optimized, including the selection of hidden units per layer. Users may alter the name and size of the layers in the neural network by using these units to control its construction. The ANN algorithm is adept at finding minima, controlling variance, and subsequently updating the model's parameters, as expressed in Equation.

$$\theta \, = \theta - \, \eta * \Delta J(\theta)`$$

Another important ANN parameter is the learning rate, which is in charge of changing the weights at each step and is essential to the model's learning process. It has to be chosen carefully since too high of a learning rate may make it difficult to identify minima, and too low of a rate can make learning go more slowly. Learning rate values that are often used are ones that are to the power of 10, such as 0.001, 0.01, 0.1, and 1. The learning rate in this model is fixed at 0.1.`

4. Results and Analysis`

The findings of the proposal are assessed in this part to provide results based on the categorization of diabetes. Performance analysis and comparison analysis are subsections of the outcome section. The PIMA and NCSU datasets are the two distinct datasets used in the performance analysis to assess the efficacy of the recommended strategy. The efficacy of the suggested strategy is compared to other approaches that have been recorded in the literature for the comparative study. The assessment criteria include f-measure, recall, accuracy, and precision.`

Analysis of PIMA Dataset performance

Performance study on the PIMA dataset involves evaluating machine learning (ML) models for diabetes prediction, exploring pre-processing, feature selection, and classification models for effectiveness. These discoveries give vital new information for healthcare applications and better the diagnosis and management of diabetes.`

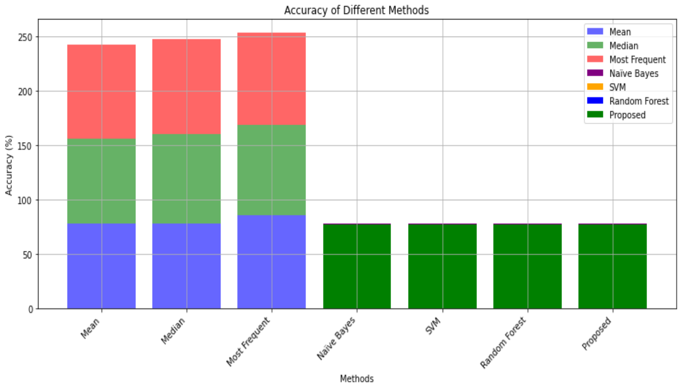
Performance analysis of PIMA Dataset for Pre-processing`

Various classifiers, such as Support Vector Machine (SVM), Random Forest, KNN, and Naïve Bayes, are employed to evaluate the efficacy of the recommended technique in this section. The PIMA dataset is used for the assessment, and Tables 1 and 2 show the outcomes. Table

1 presents the outcomes of the suggested strategy on the PIMA dataset without any pre-processing methods used, while Table 2 presents the findings with the application of pre-processing techniques. A graphical depiction of the performance analysis for the PIMA dataset is also shown in Figure 2.`

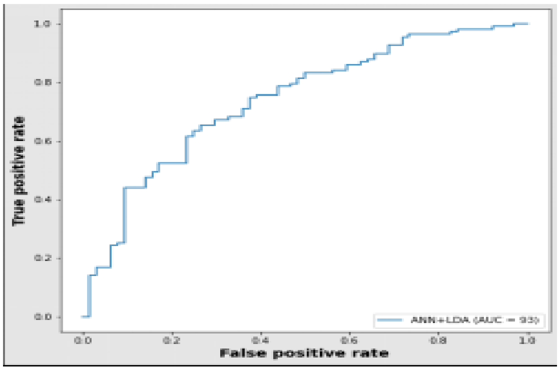
Methods	Accuracy (%)		
	Mean	Median	Most Frequent
Naïve Bayes	78.21	78.03	85.52
SVM	82.30	83.25	86.35
Random Forest	87.32	84.54	78.21
Proposed	74.16	68.23	76.98

``



Graphical representation of PIMA dataset for without pre-processing techniques`

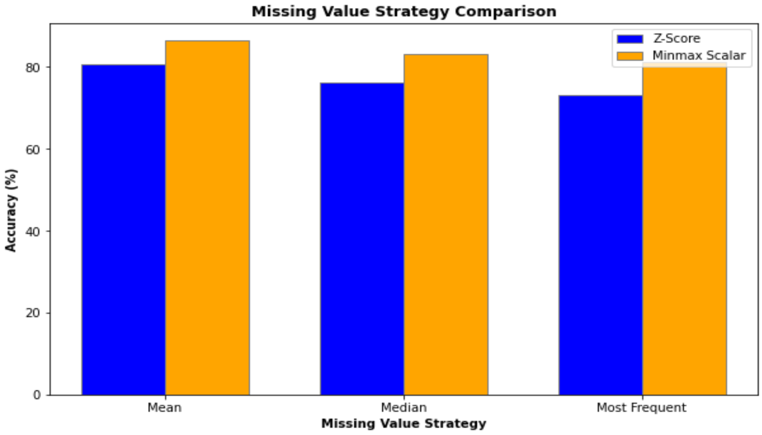
Furthermore, The performance of the suggested classifier is compared with the existing ones through an evaluation that makes use of the NSCU dataset. After contrasting the suggested approach with the state-of-the-art classification methods, results are evaluated on the NSCU dataset. `



ROC graph of accuracy for PIMA dataset`

Performance analysis for after pre-processing techniques

Missing value strategy	Z- Score	Minmax Scalar
Mean	80.65	86.45
Median	76.19	83.1
Most Frequent	73.15	81.26



Analyzing performance of Missing values

Performance analysis of PIMA Dataset for feature selection

Classifier	**Accuracy for Testing (%)**	**Accuracy for Validation (%)**
SVM	70.37	87.41
Random Forest	74.23	81.89
Correlated function	78.25	83.21

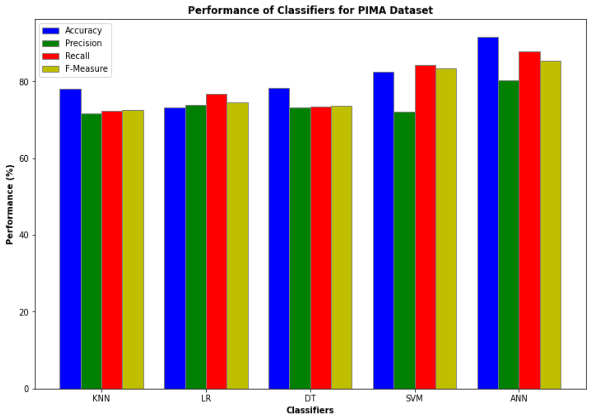
Performance analysis of Feature selection for PIMA Dataset

The results of the suggested strategy applied to the PIMA dataset using different feature selection strategies are shown in Table 3. The training and test sets of the dataset are split into proportions of 70% and 30%, respectively. The testing set receives thirty percent of the data, while the training set receives the remaining seventy percent, which is chosen at random. After experimenting with several ratios, this particular split ratio was identified as effective in producing superior outcomes.`

Performance analysis of PIMA Dataset for Classification

The performance evaluation of mentioned classifiers was conducted using the PIMA dataset, as shown in Table,

Classifiers	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
KNN	78.11	71.61	72.32	72.64
LR	73.25	73.84	76.74	74.45
DT	78.25	73.16	73.45	73.73
SVM	82.59	72.18	84.27	83.38
ANN	91.66	80.32	87.73	85.35



The table illustrates how well the suggested Artificial Neural Network (ANN) classifier performed in identifying diabetes patients from the PIMA dataset. The suggested categorization approach outperforms current approaches in a number of assessment parameters, demonstrating its effectiveness in diabetes prediction. What is really remarkable is the remarkable 91.66% classification accuracy that the suggested ANN manages to attain. With an accuracy of 78.11%, This classifier performs better than others, including Logistic Regression (LR) at 73.25%, Decision Tree (DT) at 78.25%, and Support Vector Machine (SVM) at 82.59%. `

Moreover, as other important measures like precision, recall, and F-measure demonstrate, the suggested ANN's improved performance goes beyond accuracy. Recall represents the percentage of true positives that the classifier accurately detected out of all real positives, On the other hand, precision shows the percentage of actual positive predictions out of all the positive predictions made by the classifier. The F-measure strikes a balance between accuracy and recall to provide a fair assessment of a classifier's performance. The suggested ANN routinely beats its competitors in each of these parameters, proving its dependability and robustness in diabetes prediction tasks. The suggested ANN's outstanding accuracy and extensive performance metrics highlight its potential as a useful tool in healthcare applications, especially in the early diagnosis and treatment of diabetes. The suggested classifier gives medical professionals a potent tool for raising diagnostic precision and bettering patient outcomes in the area of diabetes treatment by using cutting-edge machine learning approaches.

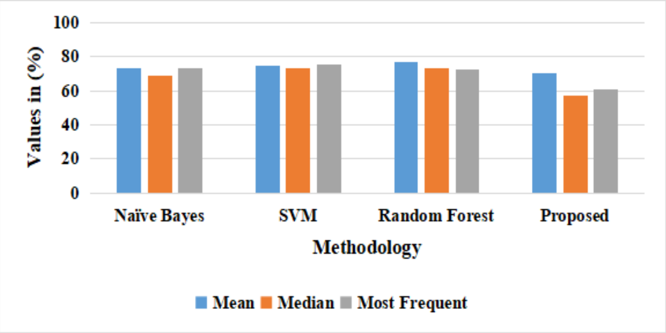
Performance Analysis of NCSU dataset

Examining the NCSU dataset's performance requires a thorough assessment of several machine learning (ML) models meant for diabetes prediction. The efficiency of several components, such as feature selection techniques, pre-processing techniques, and classification models, is examined in this investigation. Through a thorough analysis of these components, Significant advancements are made in the realm of healthcare applications, namely in the diagnosis and treatment of diabetes, thanks to the significant insights that are discovered. In order to assess machine learning models, their predictive power is put to the test using the NCSU dataset, which includes pertinent parameters including blood glucose level, age, sex, and BMI. A thorough testing and analysis process is used to assess the effectiveness of specific machine learning techniques, such as logistic regression, decision trees, support vector machines, and neural networks.Each model's recall, accuracy, precision, F1-score, and other pertinent metrics are carefully evaluated to see how well it can predict diabetes patients. Additionally, the effect of pre-processing methods on model performance is examined, including data normalization, outlier reduction, and missing value imputation.`

A thorough testing and analysis process is used to assess the effectiveness of specific machine learning techniques, such as logistic regression, decision trees, support vector machines, and neural networks.Each model's recall, accuracy, precision, F1-score, and other pertinent metrics are carefully evaluated to see how well it can predict diabetes patients. The objective of integrating these techniques into the classification process is to enhance the models' robustness and effectiveness. All things considered, the results of this performance study are a useful tool for academics, politicians, and healthcare providers. Healthcare systems may gain from better diabetes diagnosis and treatment methods by using cutting-edge ML techniques and methodologies, which will eventually enhance patient outcomes and quality of life.

Performance analysis of NCSU Dataset for Pre-processing

This section evaluates the effectiveness of the suggested method using a variety of classifiers, including Random Forest, KNN, SVM, and Naïve Bayes. Present the assessment's conclusions, which drew from NCSU datasets. These tables show the results of the suggested strategy with and without the use of pre-processing methods on the NCSU dataset. Moreover, a graphical display of the NCSU dataset's performance analysis.`



Graphical representation of the NCSU for without pre-processing techniques

Classifier	Accuracy for Testing (%)	Accuracy for Validation (%)
SVM	72.32	80.25
Random Forest	74.45	82.94
Correlated function	76.86	86.37

Performance analysis of NCSU Dataset for feature selection

Table shows that the correlation function outperforms the SVM and Random Forest classifiers in terms of training and testing accuracy after data pre-processing. Furthermore, the validation accuracy achieved by both classifiers is comparable. These findings demonstrate the associated function's greater prediction accuracy by showing a much higher true negative rate.

Performance analysis of NCSU Dataset for Classification

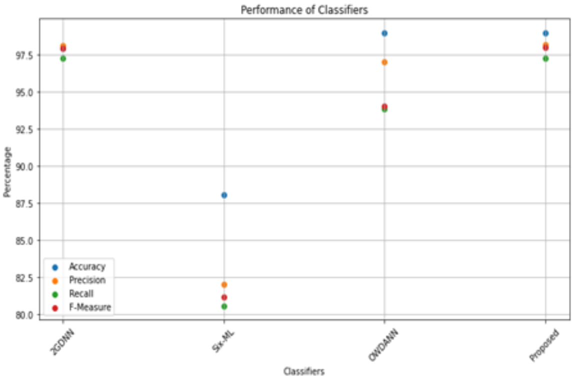
The NCSU datasets were used to evaluate the suggest classifiers' performance, as shown in the table below. Furthermore, The results of the recommended technique for the NCSU dataset are shown in the table.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
KNN	78.36	72.94	71.33	71.62
LR	72.65	74.34	75.42	73.31
DT	76.12	74.16	73.97	72.3
SVM	87.1	71.88	83.55	82.11
ANN	96.37	80.34	88.52	86.19

The table above illustrates how well the suggested ANN performs as a classifier for correctly identifying diabetes patients in the NCSU dataset. When compared to current classification techniques, the suggested classification strategy produces better overall metrics outcomes. Notably, the suggested ANN's classification accuracy of 96.37% is much greater than that of the current classifiers, including KNN's 78.36%, LR's 72.65%, DT's 76.12%, and SVM's 87.10%.

Comparative analysis

The practice of comparing data to identify trends and differences that may be utilized to gain insightful knowledge or influence choices is known as comparative analysis. This paragraph assesses the categorization strategy's performance in comparison to the latest methodologies mentioned in relevant papers. Performance indicators including F-measure score, recall, accuracy, and precision are used to evaluate candidates. Table displays the findings from the assessment of the suggested methodology for the PIMA dataset.



Graphical representation of comparative analysis for PIMA dataset

5. Conclusion`

We have employed a range of datasets and machine learning algorithms to perform a comprehensive analysis and evaluation of many strategies for diabetes prediction in this research. In an attempt to boost healthcare diagnosis and treatment outcomes, we have upgraded diabetes prediction systems via our research. We began our study by acquiring data from publicly available sources, including the NCSU and PIMA datasets. These datasets provided relevant data on features of diabetes, such as physiological testing, medical history, and demographic information. We did a variety of experiments using these datasets to analyze the efficacy of different feature selection procedures, pre-processing approaches, and classification algorithms.To assure the correctness and consistency of the data, we applied pre-processing procedures including outlier identification, missing value filling, and normalization. These processes were necessary for reducing any biases and inaccuracies in the data and preparing the datasets for analysis and classification. Next, in order to discover the most important factors for diabetes prediction, We looked at many feature selection techniques, including bivariate filtering and Pearson correlation analysis. Our purpose was to increase the usefulness and accuracy of our classification models by picking characteristics that are beneficial and decreasing the number of dimensions.`

We employed a variety of machine learning techniques, such as ensemble methods like random forests, support vector machines (SVM), k-Nearest Neighbors (KNN), decision trees (DT), logistic regression (LR), artificial neural networks (ANN), and random forests, for classification. To find out how successfully these algorithms predicted diabetes, they were trained and assessed using measures like accuracy, precision, recall, and F-measure. Our results reveal that the offered techniques perform promisingly in the prediction of diabetes, particularly those that make use of ANN and ensemble methodologies. These algorithms effectively identified diabetic people from the datasets with remarkable accuracy rates. Additionally, our research revealed the utility of preprocessing approaches in boosting data quality and the significance of feature selection in enhancing classification outcomes. To sum up, our study contributes to the continued efforts to construct trustworthy and accurate diabetes prediction models. Our objective is to support healthcare providers in the early detection and treatment of diabetes by utilizing an array of datasets and machine learning techniques. This will ultimately lead to an increase in patient outcomes and quality of life. Subsequent research can focus on refining classification algorithms, incorporating extra data sources, and employing prediction models in real clinical situations to test their usefulness and practicality.

6. Future Scope`

Even though our study's use of machine learning techniques to predict diabetes has greatly advanced the field, there are still many prospects for research and development. One path for future study might be to extract more subtle patterns and correlations from medical data by using advanced deep learning architectures like recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Further biological data sources, including genetics, lifestyle factors, and environmental exposures, may further improve prediction accuracy and provide more individualized insights into the risk of diabetes. The adoption and use of these tools for early diagnosis and intervention should also be improved by the integration of prediction models into clinical practice environments and the development of user-friendly interfaces for medical professionals. Researching the use of transfer learning techniques, which involve transferring knowledge from one domain or dataset to another, may also improve the generalization and adaptability of models to a variety of patient populations. Ultimately, ongoing research endeavors ought to focus on consistently evaluating and enhancing prediction models via the utilization of actual patient data, guaranteeing their reliability, scalability, and effectiveness in assisting clinical decision-making procedures. By doing these actions, we can help those affected by this chronic condition get better treatment and advance the field of diabetes prediction research.

References*:*

[1] Chang, V., Bailey, J., Xu, Q.A. and Sun, Z., 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, pp.1-17.

[2] aiswal, S. and Gupta, P., 2023. Diabetes Prediction Using Bi-directional Long Short-Term Memory. SN Computer Science, 4(4), p.373.

[3] Yuan, Z., Ding, H., Chao, G., Song, M., Wang, L., Ding, W. and Chu, D., 2023. A Diabetes Prediction System Based on Incomplete Fused Data Sources. Machine Learning and Knowledge Extraction, 5(2), pp.384-399.

[4] Kaul, S. and Kumar, Y., 2020. Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. SN Computer Science, 1(6), p.322.

[5] Assegie, T.A. and Nair, P.S., 2020. The performance of different machine learning models on diabetes prediction. International journal of scientific & technology research, 9(01).

[6] Cahn, A., Shoshan, A., Sagiv, T., Yesharim, R., Goshen, R., Shalev, V. and Raz, I., 2020. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes/metabolism research and reviews, 36(2), p.e3252.

[7] Li, Y.-H.; Yeh, N.-N.; Chen, S.-J.; Chung, Y.-C. Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network. Mob. Inf. Syst. 2019, 2019, 6142839. `

[8] Nibareke, T. and Laassiri, J., 2020. Using Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data, 7, pp.1-18.`

[9] Jaiswal, V., Negi, A. and Pal, T., 2021. A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), pp.435-443.`

[10] Annamalai, R. and Nedunchelian, R., 2021. Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. Computational Intelligence and Neuroscience, 2021.`