The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

AN EFFECTIVE PRE-PROCESSING TECHNIQUES FOR DIABETES MELLITUS PREDICTION IN HEALTHCARE SYSTEMS

SOUMYA K N

Assistant Professor, Department of Artificial Intelligence and Machine Learning, Jyothy Institute of Technology, Bengaluru,560082. Email: soumya.kn16@gmail.com

RAJA PRAVEEN K N

Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, 562112, India. Email: rajapraveen.k.n@gmail.com

Abstract

Diabetes Mellitus (DM) is a persistent disease caused by elevated blood sugar levels, and if left untreated, it can result in severe health complications, such as cardiac disorders, kidney damage, and stroke. However, current Machine Learning (ML) and Deep Learning (DL) approaches face challenges in accurately predicting diabetes in patients. Furthermore, this research evaluated the proposed pre-processing technique on large datasets, incorporating outlier identification and removal, missing value imputation, and standardization to enhance the accuracy of diabetes prediction. To achieve effective diabetes classification, the researchers utilized an Artificial Neural Network (ANN) with initialized weights. Data is collected from the PIMA Dataset and the North California State University (NCSU) dataset. Next, a Bivariate filter-based feature selection is performed to identify relevant features. Furthermore, the chosen features are fed into the Pearson correlation to further refine the feature set by considering a threshold value, which selects the most effective features. The experimental results showed that the proposed approach outperforms existing methods significantly, achieved better classification accuracy of 98.99%.

Keywords: Artificial Neural Network**,** Bivariate filter**,** Diabetes mellitus, Pearson correlation, Standardization.

1. INTRODUCTION

Diabetes is a persistent health condition that poses the risk of becoming a global healthcare crisis. According to the International Diabetes Federation, there are currently 382 million people worldwide living with diabetes. Shockingly, this number is projected to double by 2035, reaching a staggering 592 million. The disease is characterized by elevated blood glucose levels, making it a significant health concern for peoples and healthcare systems worldwide [1]. Early prediction of diseases like diabetes can play a crucial role in controlling and potentially saving human lives. In pursuit of this objective, this research delves into predicting diabetes by considering various attributes associated with the disease. To achieve this, the study employs the Pima Indian Diabetes Dataset and applies a range of ML classification and ensemble techniques. These methods are utilized to accurately predict the occurrence of diabetes, contributing to early intervention and improved healthcare outcomes [2]. ML is an explicit training method used to efficiently gather knowledge by building various classification and ensemble models from large datasets. ML can be employed to develop predictive models for diabetes using this extensive data repository [3]. The application of learning models as an analysis technique helps in generating inferences that are applied to large datasets. ML has garnered significant attention, particularly in disease diagnosis and object recognition.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

In the context of disease diagnosis, it enables the development of systems that can assist physicians in identifying diseases effectively [4]. The rapid advancements in Artificial Intelligence (AI), particularly in ML and computer vision, have enabled the development of applications for automating tasks that demand intelligent behavior.

In this research, an ML model was trained to predict the progression from pre-diabetes to diabetes using Electronic Medical Records (EMR), while considering the peoples historical and current medical records. The model's development and validation were thoroughly described, with data sourced from The Health Improvement Network (THIN) database for both internal and external validation. [5]. In diabetes prediction using ML, the process involves gathering pertinent data, including medical records and lifestyle factors.

From this data, relevant features are extracted, and an ML algorithm is trained to build a predictive model. This model is subsequently employed to analyze new patient data, evaluating the risk of diabetes development based on input variables [6]. By adopting this approach, the system is equipped to deliver crucial assistance in diabetes management, combining software engineering technology with ML.

The ultimate goal of the diabetes management system is to determine the optimal nutrition requirements for patients and provide meal recommendations to meet those needs. Additionally, it ensures patients are notified promptly to take their medication on time, enhancing overall healthcare management and support [7]. ML can help healthcare professionals in early detection and intervention for high-risk peoples, leading to personalized treatment plans and lifestyle adjustments. It is also applicable in developing decision support systems for diabetes management, offering real-time insights and recommendations.

Furthermore, these predictive models are valuable for population health studies, allowing the identification of risk factors and implementation of preventive measures on a larger scale [8].

Advantages of diabetes prediction using ML include early detection, personalized risk assessment, improved treatment planning, reduced healthcare costs, better patient management, enhanced quality of life, and potential prevention of complications. ML algorithms can analyze large amounts of data, identify patterns, and provide accurate predictions, leading to timely interventions and improved outcomes for peoples at risk of developing diabetes [9]. However, disadvantages of diabetes prediction using ML include the potential for false positives/negatives, reliance on accurate data input.

It has limited interpretability of the prediction models, and the risk of over-reliance on technology without proper clinical judgment. Addressing these challenges requires careful planning, stakeholder engagement, and a seamless transition strategy to effectively harness the benefits of ML while minimizing workflow disruptions [10].

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

This article uses Outlier identification and removal, filling missing values and Standardization model for improving the performance of diabetes and the major contributions are listed below:

- This research introduced a pre-processing approach involving outlier identification, missing value filling, and standardization to enhance diabetes prediction accuracy.
- The Bivariate filter method is employed for feature selection to select relevant features. Subsequently, the selected features are used as input for the Pearson correlation, which further refines the feature set by choosing the features in specified threshold value.

The rest of the article is organized in the following manner: The Section 2 describes the related works of this research and Section 3 describes the proposed methodology of this research. The Section 4 and Section 5 describes the results and conclusion of this overall research.

2. LITERATURE SURVEY

The purpose of a literature survey has to provide a critical evaluation of the existing knowledge and research on a particular subject. It helps researchers and scholars gain a deep understanding of the current state of knowledge, identify research gaps, and determine the scope and direction of their own research that are described below,

Victor Chang *et al.* [11] have developed an e-diagnosis system utilizing ML algorithms, designed for implementation on the Internet of Medical Things (IoMT) to diagnose diabetes. The proposed method was trained with three ML algorithms and evaluated to determine its ability to predict positive diabetes mellitus diagnoses based on eight specific attributes. By making IoMT and ML models accessible, healthcare professionals can be used in early detection and diagnosis, benefiting from predictive tools that enhance decision-making efficiency and correctness. However, need to further enhance diabetes mellitus prediction and address other non-communicable diseases, and required to develop novel automation and automated processes in the IoMT framework.

Chollette C *et al.* [12] introduced Twice-Growth Deep Neural Network (2GDNN) to enhance the accuracy of diabetes prediction using the PIMA Indian and LMCH datasets. Their framework involved incorporating data preprocessing techniques, Spearman correlation, and polynomial regression to bolster performance. The 2GDNN models were thoroughly evaluated with and without optimization, and the results were compared to identify the best-fitting model for the specific problem. However, at an instance, 2GDNN failed to provide a correct diagnosis, leading to the modest superiority of the ORF model in accurately handling the severity of diabetes.

Hafsa Binte Kibria *et al.* [13] suggested a diagnostic approach for diabetes using six ML algorithms and an ensemble classifier. To enhance interpretability, they provided global and local explanations for each model using Shapley Additive explanations (SHAP), presented through diverse graphs to help physicians in comprehending the model predictions effectively. This approach has the potential to significantly improve the clinical understanding of diabetes diagnosis, enabling timely interventions during the early stages of the disease. However, the

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

suggested method changed the feature scale led to varying coefficients, rendering the magnitude coefficient an unsuitable choice for determining feature importance in the model.

Isfafuzzaman Tasin *et al.* [14] implemented a six ML approaches, including decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to achieve the most accurate diabetes prediction. It also utilized a semi-supervised model with extreme gradient boosting to predict insulin features from a private dataset. To address the issue of class imbalance, SMOTE and ADASYN techniques was implemented. Moreover, it developed a user-friendly Android smartphone application and a suggested framework that allows users to input diverse features for instantaneous diabetes prediction. However, the suggested method, need to further enhance by incorporating fuzzy logic techniques and optimization approaches into the ML models.

Annamalai R and Nedunchelian [15] have developed an Optimal Weighted based Deep Artificial Neural Network (OWDANN) algorithm, for predicting diabetes mellitus disease and estimating its severity level. The system comprises two distinct phases: disease prediction and severity level estimation. In the disease prediction phase, the Pima dataset undergoes preprocessing to enhance data quality. Subsequently, relevant features were extracted from the preprocessed data, and the classification step employs the OWDANN algorithm. This approach effectively addresses noise and efficiently restores corrupted data, leading to improved prediction accuracy. However, OWDANN requires further modifications and need to enhance a wide range of scenarios and provide more comprehensive predictions for diabetes-related complications.

The proposed method aims to address the limitations found in existing diabetes prediction approaches. By enhancing data quality and eliminating inconsistencies, the suggested approach aims to significantly improve the accuracy and reliability of prediction models. As a result, the diabetes predictions generated are expected to be more robust and effective.

3. PROPOSED METHODOLOGY

The purpose of the ITSOA has to improve the accuracy of EEG based emotion classification. The process involved in various stages of classifying diabetes are data acquisition, pre-processing, feature selection and classification of diabetes.

The initial step involves acquiring data from a publicly available dataset, followed by a pre-processing phase to eliminate irrelevant or inappropriate features. Subsequently, a feature selection process is applied to choose relevant and non-redundant features.

Finally, an efficient classification is conducted using an ANN classifier to achieve accurate predictions. These steps aim to enhance the accuracy and reliability of the prediction models by improving data quality and removing inconsistencies, leading to more robust and effective diabetes predictions. The block diagram of the suggested method is illustrated in Figure. 1.

ISSN 1533-9211

DOI: 10.5281/zenodo.10202639

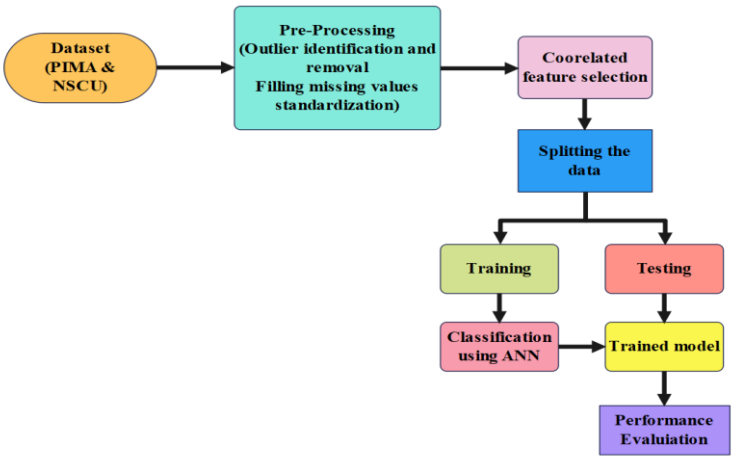

Figure 1: Flow diagram of the proposed method

3.1 Data collection

In this research, the raw data is obtained from two publicly available datasets such as PIMA Dataset [16] and North California State University (NCSU) [17] dataset. The description of the mentioned dataset is mentioned as follows:

- **PIMA:** The Pima Indian Diabetes dataset has been a standard benchmark for diabetes classification research due to its binary outcome variable, making it suitable for supervised learning, especially logistic regression. However, researchers have explored various ML algorithms to build classification models using this dataset, allowing for diversity and avoiding reliance on a single type of model.
- **NCSU:** The diabetes prediction dataset used in this study was obtained from NC State University and consists of 442 instances with 10 attributes. The feature set comprises age and sex, and for the analysis, one feature set was selected.

3.2 Data Pre-processing

After the stage of data collection, in this research, the preprocessing step of the proposed framework aims to transform the data into a processed format without complexities. It involves outlier identification and removal to eliminate extreme data points, filling missing values to ensure complete datasets, and standardization to normalize variables. The process of the proposed method is briefly outlined as follows;

3.2.1 Outlier identification and removal

The purpose of outlier identification and removal using pre-processing for diabetes prediction is to improve model accuracy by eliminating extreme data points by improving the model's performance, ensuring more reliable predictions for better outcomes. It is a crucial step in data preparation to enhance the accuracy of predictive models. Outliers are data points that significantly deviate from the majority, potentially distorting the model's performance. Through the application of pre-processing techniques, outliers can be detected and effectively

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

eliminated from the dataset. In the context of diabetes prediction, this procedure enhances the model's robustness and generalization capabilities by reducing erroneous data. The conventional approach for identifying and removing outliers in multivariate data analysis involves measuring the distance of each observation using Mahalanobis distance, as depicted in Equation (1).

$$D_M = (X - \bar{X})^T S^{-1} (X - \bar{X}) \quad (1)$$

where X is a matrix containing the average values, $x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, of the original variables. The observations associated with large values of DM are classified as outliers and then discarded. The Mahalanobis distance can be related to the principal components: it can be shown, in fact, that the sum of squares of the PC, standardized by the eigenvalue size, equals the Mahalanobis distance for observation I was illustrated in Equation (2).

$$\sum_{k=1}^{Q} \frac{z_{ik}^2}{l_k} = \frac{z_{i1}^2}{l_1} + \frac{z_{i2}^2}{l_2} + \ldots + \frac{z_{iQ}^2}{l_Q} = D_{M,i} \quad (2)$$

In high-dimensional datasets, certain outliers may not be evident when analyzing individual dimensions, making them undetectable using univariate criteria. Therefore, a multivariate approach becomes essential. In this context, Principal Component Analysis (PCA) serves as an ideal tool for identifying and eliminating outlier observations effectively.

3.2.2 Filling missing values

The purpose of filling missing values using pre-processing for diabetes prediction is to ensure that the dataset is complete and ready for analysis, as missing data can negatively impact the performance of predictive models. By imputing missing values, the model can make more accurate predictions for improved diabetes diagnosis and treatment. Missing values may arise due to diverse factors, such as data entry errors or incomplete information. To maintain the integrity and precision of the predictive model, these missing values are substituted with estimated or imputed values using various techniques.

Common methods for filling missing values involve mean, median, or mode imputation, where the central tendency of the existing data is utilized to replace the absent entries. In addition, more advanced techniques like regression imputation or K-Nearest Neighbors (KNN) imputation can be utilized, utilizing relationships between variables observed in neighboring data points. In the proposed framework, missing or null values were imputed using the mean values of the attributes rather than dropping them, as represented in equation (3). Imputation with the mean offers benefits as it fills in continuous data without introducing outliers.

$$(x) = f(x) = \begin{cases} mean(x), & if\ x = null/missed \\ x, & otherwise \end{cases} \quad (3)$$

where x is the instances of the feature vector that lies in n-dimensional space, x ∈ R.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

3.2.3 Standardization

The purpose of standardization is to normalize data, ensuring features are on a consistent scale for improved diabetes prediction. Z-score normalization, also known as standardization, is a technique used to rescale attributes to achieve a standard normal distribution with zero mean and unit variance. Standardization (R), as depicted in equation (4), also contributes to reducing the skewness of the data distribution.

$$R(x) = \frac{x - \bar{x}}{\sigma} \quad (4)$$

where x is the n-dimensional instances of the feature vector, $x \in R^n . \bar{x} \in R^n$ and $\sigma \in R^n$ are the mean and standard deviation of the attributes. Nonetheless, in various ML models, such as tree-based models, standardizing features may not necessarily guarantee significant improvements in performance.

3.3 Feature selection

The pre-processed output is obtained from the initial pre-processing stage and then fed as input to the subsequent feature selection stage. Feature selection aims to improve classification accuracy by identifying the most relevant features that facilitate the diabetes classification process. In this research, the feature selection process utilizes the Bivariate statistics approach, which efficiently selects appropriate and relevant features from large datasets like PIMA and NCSU. The Bivariate filter is employed for feature extraction, effectively integrating heterogeneous data layers to address uncertainties in the input data. Additionally, this filter utilizes a certainty factor to identify relevant features, and its evaluation is carried out using equation (5) as follows:

$$CF = \begin{cases} \frac{PP_a - PP_s}{PP_a(1-PP_s)}, if\ PP_a \geq PP_s \\ \frac{PP_a - PP_s}{PP_s(1-PP_a)}, if\ \ | PP_a < PP_s \end{cases} \quad (5)$$

Where the conditional probability of CF is denoted as PP_a and the prior probability of the selected features are represented as PP_s. The value of PP_a and PP_s is evaluated using the equation (6) and (7) respectively.

$$PP_a = P\{S|B\} \quad (6)$$

$$PP_s = P\{S\} \quad (7)$$

Where the conditional probability unit of B is represented as $P\{S|B\}$. Positive results indicate an increase in the certainty value, whereas negative results signify a decrease in the certainty value. Furthermore, the features are extracted using the Weight of Evidence (WoE) based on the Bayesian probability approach, utilizing weights to determine their significance. The positive and negative weights of the features are evaluated using two parameters, W^+ and W^-, as represented in equations (8) and (9) respectively.

$$W^+ = ln\frac{P\{B|A\}}{P\{B|\bar{A}\}} \quad (8)$$

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

$$W^{-} = ln\frac{P\{\bar{B}|A\}}{P\{\bar{B}|\bar{A}\}} \quad (9)$$

The logarithm and probability values are represented as P and ln respectively. The features selected using the bivariate filter method are then used as input for Pearson correlation, which identifies effective features based on a specified threshold value.

3.3.1 Pearson Correlation

To enhance the relationship between Pearson correlation and diabetic characteristics, the parameters are optimized to remove redundant information. The Pearson correlation coefficient is used to assess linear relationships between random variables. Equation (10) below illustrates the linear correlation between two continuous variables.

$$r_{xy} = \frac{\sum(x_i-\bar{x})\ \sum y_i-\overline{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\overline{y})^2}} \quad (10)$$

If $r_{xy} = 1$, x and y are a totally positive correlation, If $r_{xy} = 0$, the linear correlation between x and y is not obvious and when $r_{xy} = -1$, x and y are a totally negative correlation.

The Pearson correlation, with an R-value of 0.12, exhibits a less significant effect on Diabetes. The relationship between certain features and diabetes is found to be moderate (r = 0.33, r = -0.42, r = 0.23). It is important to note that correlation does not imply causation. Utilizing this information, relevant features strongly associated with diabetes are identified and selected as input variables for precise disease classification. The output of Pearson correlation is then fed into the classification process to identify cases of type II diabetes.

3.4 Classification using ANN

After the stage of feature selection, the classification was employed by both the PIMA and NCSU datasets are utilized. One of the ML algorithms employed for classification is Artificial Neural Network (ANN), known for providing more accurate results compared to existing techniques. ANN comprises one or more hidden layers, with neurons processing the information. Each node functions as an activation node, classifying the outcome of artificial neurons to achieve improved results. To minimize the training error, several parameters can be optimized, including the selection of hidden units per layer. These units determine the name and size of the layers, enabling users to customize the neural network's structure. The ANN algorithm is adept at finding minima, controlling variance, and subsequently updating the model's parameters, as expressed in Equation (11).

$$\theta = \theta - \eta * \nabla J(\theta) \quad (11)$$

The learning rate is another critical parameter in ANN, responsible for adjusting the weights at each step and playing a crucial role in the model's learning process. It must be carefully chosen, as a learning rate that is too high may hinder the selection of minima, while one that is too low can slow down the learning speed. Commonly selected learning rate values are in the power of 10, such as 0.001, 0.01, 0.1, and 1. In this model, the learning rate is set to 0.1.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

4. RESULTS AND ANALYSIS

In this section, the results obtained from the proposed is evaluated to obtain the results based on diabetes classification. The result section is sub-sectioned to performance analysis and the comparative analysis. The performance analysis involves assessing the efficiency of the proposed approach on two distinct datasets, namely PIMA and NCSU. For the comparative analysis, the proposed approach's effectiveness is evaluated against existing approaches documented in the literature. The evaluation metrics encompass accuracy, precision, recall, and f-measure, which are computed using the equations (12-15) as provided below.:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

$$F1\ measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (15)$$

Where the TP, FP, TN and FN is the True Positive, False Positive, True Negative and False Negative respectively.

4.1 Experimental setup

The proposed ANN-LDA classification approach was implemented on a system with the following specifications: Anaconda Navigator 3.5.2.0 (64-bit), Python 3.7 software, Windows 10 (64-bit) operating system, Intel Core i7 processor, and 16 GB of random-access memory.

4.2 Performance analysis of PIMA Dataset

The performance analysis of the PIMA dataset includes evaluating ML models for diabetes prediction, exploring pre-processing, Feature selection and Classification models for effectiveness, thus enabling valuable insights for healthcare applications and improving diabetes diagnosis and management.

4.2.1 Performance analysis of PIMA Dataset for Pre-processing

In this subsection, we evaluate the performance of the proposed approach using various classifiers, including Naïve Bayes, KNN, Support Vector Machine (SVM), and Random Forest. The evaluation is conducted on the PIMA dataset, and the results are presented in Table 1 and Table 2.

Table 1 displays the results obtained from the proposed method on the PIMA dataset without applying any pre-processing techniques, while Table 2 shows the results after applying pre-processing techniques. Additionally, Figure 2 provides a graphical representation of the performance analysis for the PIMA dataset.

ISSN 1533-9211

DOI: 10.5281/zenodo.10202639

Table 1: PIMA dataset for without pre-processing techniques

Methods	Accuracy (%)		
	Mean	Median	Most Frequent
Naïve Bayes	74.21	68.03	74.52
SVM	75.30	74.25	76.35
Random Forest	76.32	74.54	73.21
Proposed	70.16	58.23	61.98

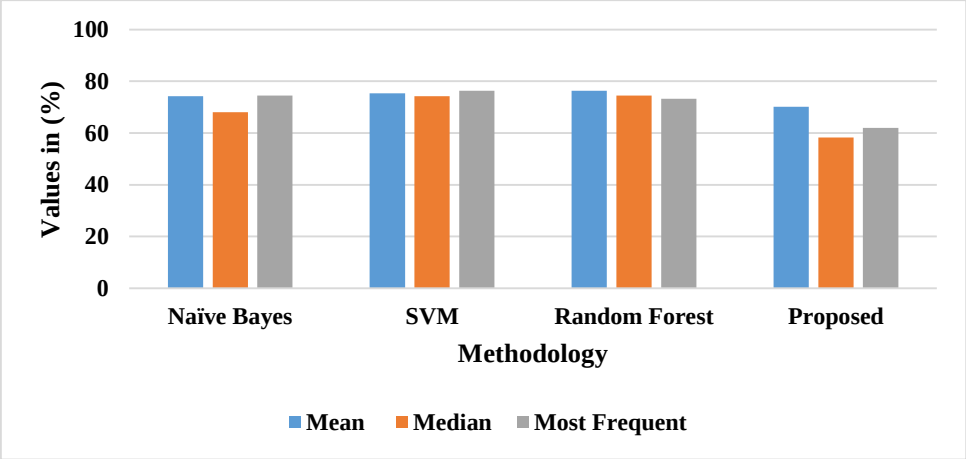

Figure 2: Graphical representation of PIMA dataset for without pre-processing techniques

Table 2: Performance analysis for after pre-processing techniques

Missing value strategy	Z- Score	Minmax Scalar
Mean	74.65	83.45
Median	60.19	81.10
Most Frequent	64.15	80.26

The results from Table 1 and Table 2 demonstrate that the proposed method serves as an excellent classifier for distinguishing diabetic patients in the PIMA dataset. The performance of the proposed classification approach outperforms existing methods in terms of overall metrics, particularly in accuracy.

4.2.2 Performance analysis of PIMA Dataset for feature selection

Table 3 presents the results obtained from the proposed method applied to the PIMA dataset using various feature selection techniques. The dataset is divided into training and test sets in a ratio of 70% and 30%, respectively. The training set consists of 70% of the data randomly chosen, while the remaining 30% is allocated to the testing set. This specific split ratio was determined after exploring various combinations and has proven to be efficient in achieving better results.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

Table 3: Performance analysis of Feature selection for PIMA Dataset

Classifier	Accuracy for Testing (%)	Accuracy for Validation (%)
SVM	75.37	81.41
Random Forest	77.23	82.89
Correlated function	78.25	85.21

According to the data presented in Table 3, the correlated function outperforms the SVM and Random Forest classifiers in terms of training and testing accuracy, following data pre-processing. Furthermore, both classifiers achieve similar validation accuracy. Notably, the correlated function demonstrates a substantially higher true negative rate, implying a more accurate prediction capability.

4.2.3 Performance analysis of PIMA Dataset for Classification

The performance evaluation of mentioned classifiers was conducted using the PIMA dataset, as shown in Table 4. Additionally, Table 4 displays the results obtained from the proposed approach for the same PIMA dataset.

Table 4: Comparing the performance of the classifiers for PIMA dataset

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
KNN	75.11	71.61	72.32	72.64
LR	76.25	73.84	76.74	74.45
DT	75.25	73.16	73.45	73.73
SVM	86.59	72.18	84.27	83.38
ANN	90.66	80.32	87.73	85.35

Table 4 demonstrates that the proposed ANN serves as a highly effective classifier for distinguishing diabetic patients within the PIMA dataset. The proposed classification approach outperforms existing methods across various metrics. Additionally, the classification accuracy of the proposed ANN reaches an impressive 90.66%, surpassing the accuracies of other classifiers, such as KNN of 75.11%, LR of 76.25%, DT of 75.25%, and SVM of 86.59% respectively.

4.3 Performance Analysis of NCSU dataset

The performance analysis of the NCSU dataset involves assessing various ML models for predicting diabetes. It also encompasses exploring the effectiveness of pre-processing, feature selection, and classification models. These findings provide valuable insights for healthcare applications, leading to enhancements in diabetes diagnosis and management.

4.3.1 Performance analysis of NCSU Dataset for Pre-processing

Within this sub-section, the proposed approach's performance is assessed using various classifiers, including Naïve Bayes, KNN, SVM, and Random Forest. The evaluation is based on the NCSU datasets, and the results are presented in Table 5 and Table 6. These tables illustrate the outcomes of the proposed method on the NCSU dataset, both without and after employing pre-processing techniques. Furthermore, a graphical representation of the performance analysis for the NCSU dataset is provided in Figure 3.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

Table 5: NCSU dataset for without pre-processing techniques

Methods	Accuracy (%)		
	Mean	Median	Most Frequent
Naïve Bayes	73.31	69.06	73.11
SVM	74.36	73.35	75.22
Random Forest	76.66	73.36	72.35
Proposed	70.13	57.22	60.46

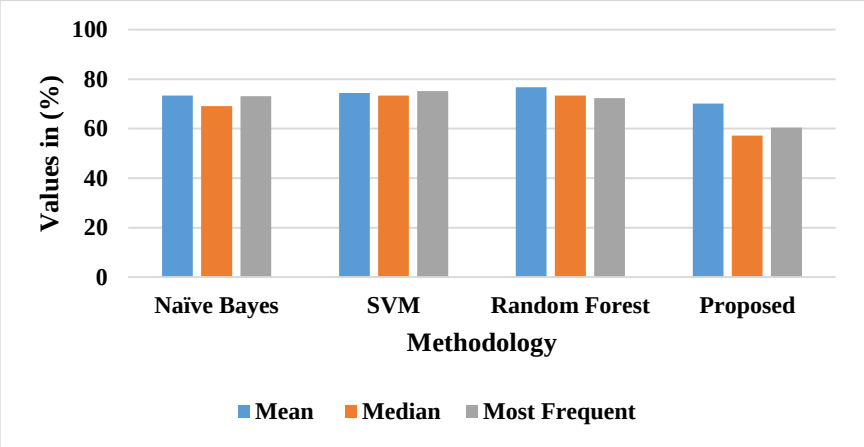

Figure 3: Graphical representation of the NCSU for without pre-processing techniques

Table 6: Performance analysis for NCSU dataset after pre-processing techniques

Missing value strategy	Z- Score	Minmax Scalar
Mean	73.21	82.47
Median	60.32	80.65
Most Frequent	63.51	80.51

The results from Table 5 and Table 6 demonstrate that the proposed method serves as an outstanding classifier in accurately identifying diabetic patients within the NCSU dataset. When compared with existing classification methods, the proposed approach achieves superior results in overall metrics, particularly in terms of accuracy.

4.3.2 Performance analysis of NCSU Dataset for feature selection

Table 7 presents the results obtained from the proposed method applied to the NCSU dataset using various feature selection techniques. The dataset is divided into training and test sets at a ratio of 70% and 30%, respectively. This split is determined after exploring various combinations, proving its efficiency in achieving optimal performance.

Table 7: Performance analysis of Feature selection for NCSU Dataset

Classifier	Accuracy for Testing (%)	Accuracy for Validation (%)
SVM	74.32	80.25
Random Forest	76.45	81.94
Correlated function	78.86	85.37

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

Table 7 reveals that after data pre-processing, the training accuracy and testing accuracy of the correlated function surpass those of the SVM and Random Forest classifiers. Additionally, both classifiers achieve similar validation accuracy. These results indicate that the correlated function exhibits a significantly higher true negative rate, highlighting its superior correctness in predictions.

4.3.3 Performance analysis of NCSU Dataset for Classification

The performance evaluation of the recommend classifiers was conducted using the NCSU datasets, as depicted in Table 8. Additionally, Table 8 presents the results obtained from the proposed approach for the NCSU dataset.

Table 8: Comparing the performance of the classifiers for NCSU dataset

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
KNN	74.36	72.94	71.33	71.62
LR	75.65	74.34	75.42	73.31
DT	74.12	74.16	73.97	72.30
SVM	85.10	71.88	83.55	82.11
ANN	90.37	80.34	88.52	86.19

Table 4 demonstrates that the proposed ANN serves as an outstanding classifier for accurately classifying diabetic patients within the NCSU dataset. The proposed classification approach achieves superior results in overall metrics compared to existing classification methods. Notably, the classification accuracy of the proposed ANN reaches 90.37%, which is significantly higher than the accuracies of existing classifiers, such as KNN of 74.36%, LR of 75.65%, DT of 74.12%, and SVM of 85.10%.

4.4 Comparative analysis

Comparative analysis refers to the Comparison of data to identify similarities and differences for meaningful insights or decision-making. In this subsection, the classification approach's performance is assessed by comparing it with existing approaches listed in related works. Evaluation is based on performance metrics such as accuracy, precision, recall, and F-measure score. The results obtained from evaluating the proposed approach for the PIMA dataset are presented in Table 9.

Table 9: Comparative analysis of various classifier for PIMA dataset

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
2GDNN [12]	97.93	98.11	97.23	97.95
Six- ML [14]	88.05	82.02	80.56	81.21
OWDANN [15]	98.97	97.02	93.84	94.04
Proposed	**98.99**	**98.15**	**97.25**	**97.96**

The graphical representation of the comparative analysis for PIMA dataset was illustrated in Figure 4.

ISSN 1533-9211

DOI: 10.5281/zenodo.10202639

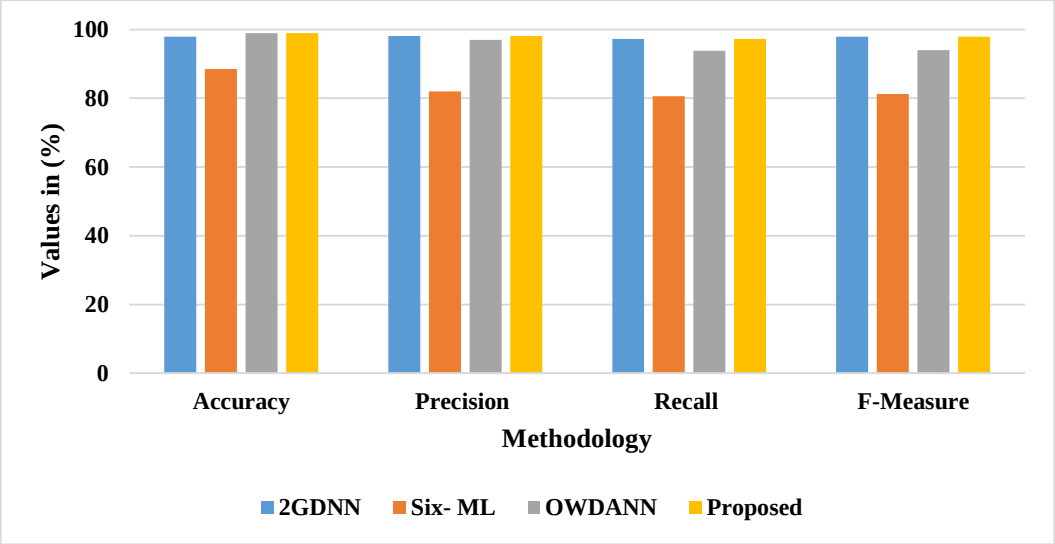

Figure 4: Graphical representation of comparative analysis for PIMA dataset

Table 9 and Figure 4 demonstrate that the proposed classification approach outperformed other methods in overall performance metrics. The accuracy achieved by the proposed approach is 98.99%, significantly higher than the Twice Growth Deep Neural Network (2GDNN) (97.93%), Six ML methods (88.05%), and Optimal Weighted based Deep Artificial Neural Network (OWDANN) (98.97%).

5. CONCLUSION

The research introduces a pre-processing approach involving outlier identification, missing value filling, and standardization to enhance diabetes Mellitus prediction accuracy. The proposed method utilizes an ANN with optimized weight initialization for effective diabetes classification. The approach's performance is evaluated on both PIMA and NCSU datasets using accuracy, precision, recall, and F-measure metrics. Following the Bivariate filter-based feature selection stage, relevant features are selected, and the chosen features undergo Pearson correlation analysis using a threshold value. The resulting effective features are then utilized as input for the ANN classifier, performing the final classification. The proposed approach outperforms existing methods in overall metrics, with an accuracy of 98.99%, surpassing 2GDNN, Six ML methods, and OWDANN of 7.93%, 88.05%, and 98.97% respectively. Future work can explore incorporating meta-heuristic algorithms to further enhance accuracy by selecting appropriate features.

References

1) Rani, K.J., 2020. Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *6*, pp.294-305.

2) Soni, M. and Varma, S., 2020. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert) Volume*, *9*.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10202639

3) 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.

4) 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).

5) 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.

6) Sowah, R.A., Bampoe-Addo, A.A., Armoo, S.K., Saalia, F.K., Gatsi, F. and Sarkodie-Mensah, B., 2020. Design and development of diabetes management system using machine learning. *International journal of telemedicine and applications*, *2020*.

7) Vehí, J., Contreras, I., Oviedo, S., Biagi, L. and Bertachi, A., 2020. Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. *Health informatics journal*, *26*(1), pp.703-718.

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) Ramesh, J., Aburukba, R. and Sagahyroon, A., 2021. A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), pp.45-57.

11) 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.

12) Olisah, C.C., Smith, L. and Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, p.106773.

13) Kibria, H.B., Nahiduzzaman, M., Goni, M.O.F., Ahsan, M. and Haider, J., 2022. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors*, *22*(19), p.7268.

14) Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), pp.1-10.

15) Annamalai, R. and Nedunchelian, R., 2021. Diabetes mellitus prediction and severity level estimation using OWDANN algorithm. *Computational Intelligence and Neuroscience*, *2021*.

16) 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.

17) SVKR Rajeswari, V., 2021. Prediction of diabetes mellitus using machine learning algorithm. *Annals of the Romanian Society for Cell Biology*, pp.5655-5662.

18) Parente, A. and Sutherland, J.C., 2013. Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity. *Combustion and flame*, *160*(2), pp.340-350.

19) Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, pp.76516-76531.

20) Naz, H. and Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*, pp.391-403.