Journal of Engineering Technology and Applied Physics

Predicting Diabetes Mellitus with Machine Learning Techniques

Tong Hau Lee*, Ng Hu and Harannesh Arul Ananthan
Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia.
**Corresponding author:* hltong@mmu.edu.my, *ORCiD:* 0000-0002-3128-585X
https://doi.org/10.33093/jetap.2024.6.1.12
Manuscript Received: 6 October 2023, Accepted: 20 December 2023, Published: 15 March 2024

***Abstract* — This study addresses the challenge of accurately identifying diabetes mellitus in individuals. Utilizing accessible online and real-world diagnostic data, we employ machine learning models, including Support Vector Machine, Random Forest, Naïve Bayes, eXtreme Gradient Boosting, and Deep Neural Network, on the PIMA Indian Diabetes and NHANES 1999-2016 datasets. Rigorous data pre-processing steps were conducted, handling null values, outliers, and imbalanced data together with data normalization. Our results reveal that the RF model achieves a 79% accuracy for binary classification on the PIMA Indian Diabetes dataset, using a 60:40 train-test split with BORUTA selected features. Meanwhile, the XGBoost model excels on the NHANES 1999-2016 dataset, achieving 92% accuracy for binary and 91% for multiclass classification respectively.**

Keywords—Diabetes Mellitus, Machine Learning, Accuracy

I. INTRODUCTION

Diabetes Mellitus (DM) poses a significant global health challenge, affecting millions and exhibiting a rising prevalence, leading to severe health consequences [1]. DM encompasses Type 1, Type 2, and Gestational Diabetes, each with distinct characteristics and impacts, demanding early detection and effective management [2].

Utilizing Machine Learning (ML) techniques, the purpose of this research is to contribute to DM prediction based on patient medical data, ultimately advancing early intervention and patient care. The study addresses crucial questions, including feature relevance, model selection, and appropriate evaluation metrics. Expected outcomes encompass the identification of critical predictive features, a comparative assessment of various ML methods, validation of model performance using real-world datasets like PIMA Indian Diabetes and National Health and Nutrition Examination Survey (NHANES) 1999-2016 datasets, and insights into factors impacting DM prediction.

The project's scope involves an in-depth analysis of ML techniques and models, with a focus on feature relevance, model selection, and performance evaluation. This research will utilize real-world datasets, including PIMA Indian Diabetes and NHANES 1999-2016, to validate and refine the predictive models. This research is motivated by the pressing need to enhance DM prevention, management, and patient outcomes. This research aims to develop accurate prediction models for the benefit of individuals and healthcare professionals alike.

This research intended to find the specific attributes in the patient's medical data hold greater significance in predicting DM. Besides, among the diverse array of ML models presently accessible, pinpoint the suitable models for predicting DM compared to alternative existing approaches.

II. LITERATURE REVIEW

A comprehensive literature review was conducted to augment the project's foundational knowledge. This review encompassed two critical aspects: first, understanding the global background and trends of DM, as discussed in the previous chapter; and second, examining prior research by various scholars involving DM prediction through ML techniques. This chapter provides a detailed account of the insights and findings derived from this extensive review.

A. Datasets Used

Previous research on predicting DM using ML has leveraged various datasets. The widely adopted dataset in this domain is the PIMA Indian Diabetes Dataset, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and accessible through Kaggle.

This dataset comprises records from 768 individuals, including 268 with DM, characterized by 8 medical features and an output feature indicating DM presence.

Researchers have also explored the National Health and Nutrition Examination Survey (NHANES) dataset, spanning 1999 to 2016, which encompasses comprehensive health and nutritional data for adults and children in the United States. NHANES offers 51 health-related features, including demographic, questionnaire, examination, and laboratory data, covering 37079 individuals, among whom 708 had borderline DM, 4144 had DM, and 32227 did not have DM [3, 4].

Additionally, real-world diagnostic data from the Medical Centre of Chittagong (MCC) in Bangladesh has been utilized in DM prediction research [5]. This dataset comprises multiple attributes related to DM from 200 patients.

Furthermore, a dataset comprising symptoms of DM, obtainable without medical examination, consisting of 521 records from Kaggle, has been employed in research [6].

These datasets serve as valuable resources for training and evaluating ML models to predict DM based on diverse patient characteristics and medical data.

B. Data Preprocessing Methods Used

Data preprocessing is a critical step in DM prediction using ML techniques. This section highlights the key data preprocessing methods employed in prior research.

i) *Handling Missing Values:* Handling missing values is crucial in data preprocessing. Researchers have used various techniques to manage missing data. For instance, some replaced missing values with the mean or median of the corresponding attribute, while others employed standard deviation values or feature class means for imputation [7-10].
ii) *Data Scaling and Normalization:* Data scaling and normalization techniques have been applied to ensure that features are on a consistent scale for analysis. Techniques like subtracting the mean and dividing it by the variance or using min-max scaling have been employed [4, 8, 10].
iii) *Feature Selection:* Feature selection aims to eliminate redundant features and improve model performance. Methods such as correlation matrix analysis, random forest feature importance, Recursive Feature Elimination with Random Forest Importance, and Analysis of Variance (ANOVA) have been used for feature selection [4, 7, 11].
iv) *Handling Imbalanced Datasets:* Addressing imbalanced datasets is crucial to prevent bias in model training. Researchers have employed oversampling techniques and class weight assignment to balance datasets and enhance predictive model accuracy [5, 10].
v) *Training & Testing Data Splitting:* Datasets have been split into training and testing sets for model training and evaluation. Various train-test split ratios have been used to ensure unbiased model performance assessment [9].

C. Machine Learning Models Used

i) *Support Vector Machine:* The Support Vector Machine (SVM) is an extensively used supervised ML algorithm for DM prediction. SVM excels in both classification and regression tasks, offering flexibility and robustness. SVM's core concept involves finding a hyperplane in an N-dimensional space to separate data points effectively. Different kernels, such as Linear and Radial Basis Function (RBF), enable SVM to handle various data distributions. Researchers have highlighted SVM's capability to manage high-dimensional spaces, create hyperplanes with maximum margins, and classify diabetes cases accurately [5, 8, 10-12].
ii) *Random Forest:* Random Forest (RF) is a versatile supervised ML algorithm employed for DM prediction. RF employs ensemble learning, combining multiple decision trees to enhance predictive accuracy and reduce overfitting. RF's injection of randomness through random sampling from training data and features improves model stability. Researchers have emphasized RF's effectiveness in handling both continuous and categorical variables, making it suitable for diverse datasets [8, 10-13].
iii) *Naïve Bayes:* Naïve Bayes (NB) is a popular probabilistic classification technique for DM prediction. Despite assuming feature independence, NB remains a well-known and effective classifier. It computes probabilistic results by combining values from the dataset. NB's simplicity, understandability, and speed make it suitable for large datasets and various applications [5, 8, 13].
iv) *Neural Network:* Neural Network (NN) models, including deep learning networks, are widely used for DM classification. NNs simulate the human brain's learning process and can model complex patterns in data. Artificial Neural Networks (ANNs), a type of NN, employ interconnected artificial neurons with weighted connections to classify input information. NNs have demonstrated their effectiveness in various domains, including DM classification [4, 8, 13].
v) *eXtreme Gradient Boosting:* eXtreme Gradient Boosting (XGBoost) is a powerful implementation of gradient boosted trees. It excels in supervised ML tasks, including DM prediction, by integrating predictions from multiple weak models. XGBoost's iterative learning process improves prediction accuracy, and its scalability enables efficient handling of large datasets. The algorithm's robustness and

versatility, combined with the ability to fine-tune hyperparameters, make it a popular choice in ML competitions and applications [14].

In summary, these machine learning algorithms play a crucial role in predicting DM, each offering unique strengths and capabilities. Understanding their fundamentals and applications is essential for effective DM prediction.

III. RESEARCH METHODOLOGY

In this section, the methodology of the research is shown in Fig. 1.

A. Data Collection and Validation

For this study, we collected and validated data from two datasets:

i) *PIMA Indian Diabetes Dataset:* The PIMA Indian Diabetes dataset, obtained from Kaggle, contains 9 features, with 8 being medical predictor variables for diagnosing the incidence of DM, and 1 indicating the presence of DM, including records of 768 patients. The features in this dataset along with their descriptions can be found in Table I.

Table I: Description of features in PIMA Indian Diabetes dataset.

No.	Attribute Names	Attribute Description	Data Types	Sample Values
01	Pregnancies	Number of times pregnant	Numeric	3, 5, 7
02	Glucose	Glucose concentration level (mg/dl)	Numeric	85, 148, 188
03	Blood Pressure	Diastolic blood pressure (mmHg)	Numeric	65, 83, 97
04	Skin Thickness	Triceps skin fold thickness (mm)	Numeric	15, 25, 30
05	Insulin	2-hour serum insulin (mm U/ml)	Numeric	0, 88, 122
06	BMI	Body Mass Index (Kg/m^2)	Numeric	25.5, 28.9, 33.7
07	Diabetes Pedigree Function	Pedigree utility of diabetes	Numeric	0.457, 0.758, 0.936
08	Age	Age in years	Numeric	24, 29, 33
09	Outcome	Presence of diabetes	Numeric	0 (Indicates no diabetes) and 1 (Indicates diabetes)

ii) *NHANES 1999-2016 Dataset:* The NHANES 1999-2016 dataset contains demographic, examination, laboratory, and questionnaire data of U.S. patients. It consists of 51 features, with 1 indicating the presence of DM. This dataset contains 37079 records. The features in this dataset have been listed down in Table II.

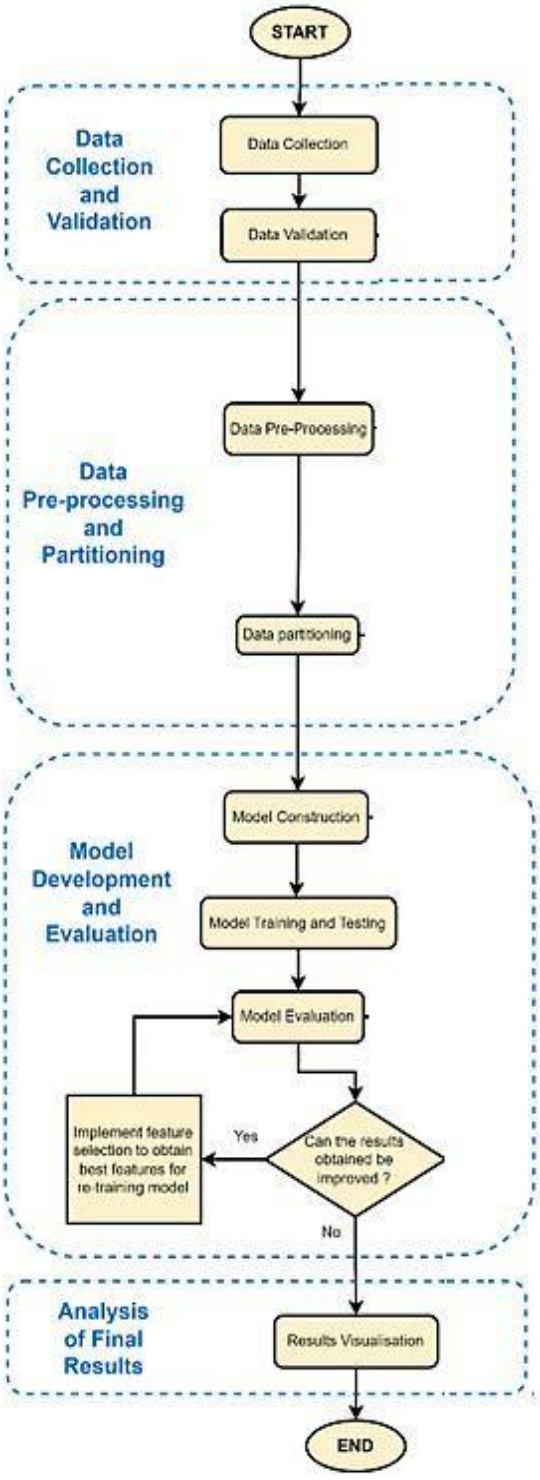

Fig. 1. Research Framework.

Table II: List of features in NHANES 1999-2016 dataset.

No.	Attribute Names	Data Types	Sample Values
01	SEQN	Numeric	2, 5, 12
02	Gender	Numeric	1 - Male, 2 - Female
03	Age	Numeric	37, 49, 77
04	Annual-Family-Income	Numeric	3, 5, 8
05	Ratio-Family-Income-Poverty	Numeric	2.67, 1.07, 4.93
06	X60-sec-pulse	Numeric	64, 72, 102

07	Systolic	Numeric	98, 122, 174
08	Diastolic	Numeric	56, 66, 99
09	Weight	Numeric	63.6, 75.4, 92.5
10	Height	Numeric	157.7, 166.2, 178.3
11	Body-Mass-Index	Numeric	24.90, 27.33, 30.62
12	White-Blood-Cells	Numeric	5.9, 9.1, 11.6
13	Lymphocyte	Numeric	13.1, 21.1, 29.8
14	Monocyte	Numeric	3.8, 6.2, 9.0
15	Eosinophils	Numeric	1.7, 3.2, 4.4
16	Basophils	Numeric	0.4, 0.5, 0.6
17	Red-Blood-Cells	Numeric	4.73, 5.13, 5.76
18	Hemoglobin	Numeric	14.1, 16.0, 16.8
19	Mean-Cell-Vol	Numeric	83.5, 88.5, 91.1
20	Mean-Cell-Hgb-Conc.	Numeric	27.8, 29.3, 31.3
21	Mean-cell-Hemoglobin	Numeric	33.3, 33.6, 34.5
22	Platelet-count	Numeric	160.0, 209.0, 357.0
23	Mean-Platelet-Vol	Numeric	7.7, 8.8, 10.4
24	Segmented-Neutrophils	Numeric	52.2, 63.7, 82.4
25	Hematocrit	Numeric	41.8, 43.6, 50.4
26	Red-Cell-Distribution-Width	Numeric	12.4, 13.7, 14.4
27	Albumin	Numeric	40, 45, 47
28	ALP	Numeric	63, 103, 110
29	AST	Numeric	17, 22, 24
30	ALT	Numeric	16, 28, 35
31	Cholesterol	Numeric	4.42, 5.25, 7.94
32	Creatinine	Numeric	61.9, 70.7
33	Glucose	Numeric	4.330, 6.384, 7.882
34	GGT	Numeric	20, 24, 32
35	Iron	Numeric	11.82, 12.18, 24.54
36	LDH	Numeric	133, 150, 181
37	Phosphorus	Numeric	0.904, 1.033, 1.130
38	Bilirubin	Numeric	6.8, 8.6, 12.0
39	Protein	Numeric	66.0, 73.0, 79.0
40	Uric Acid	Numeric	362.8, 404.5, 410.4
41	Triglycerides	Numeric	0.756, 1.581, 3.635
42	Total-Cholesterol	Numeric	4.03, 5.56, 8.12
43	HDL	Numeric	0.98, 1.08, 1.39
44	Glycohemoglobin	Numeric	4.7, 5.8, 7.6
45	Vigorous-work	Numeric	1 - Yes, 2 - No, 3 - Unable to do activity
46	Moderate-work	Numeric	1 - Yes, 2 - No, 3 - Unable to do activity
47	Health-Insurance	Numeric	1 - Yes, 2 - No, 7 - Refused, 9 - Don't know
48	Blood-Rel-Stroke	Numeric	1 - Yes, 2 - No
49	Coronary Heart Disease	Numeric	0 - No, 1 - Yes
50	Blood-Rel-Diabetes	Numeric	1 - Yes, 2 - No
51	Diabetes	Numeric	1 - Yes, 2 - No, 3 - Borderline diabetes

B. Data Preprocessing and Partitioning

i) *Removing Unrelated Features:* Unrelated features that did not contribute to the prediction of DM were removed early on to reduce dimensionality and overfitting.

ii) *Addressing Zero/Null Values:* Zero or null values in the datasets were either replaced with their median values due to them being less sensitive to outliers.

iii) *Checking for Duplicate Values:* Duplicate values were checked for in the datasets and removed if present to avoid biased predictions from being made on the duplicate target classes.

iv) *Re-encoding Target and Categorical Features:* Categorical features, including the target feature, were re-encoded from the original datasets to make them better suited for binary and multiclass classification tasks.

v) *Handling Major Outliers:* Major outliers were identified and replaced with the median value of the respective continuous features from the datasets containing outliers using the Inter Quartile Range (IQR) method to avoid skewed model performance.

vi) *Data Normalization:* We normalized the continuous features in both datasets to ensure all the continuous features in both datasets were on a uniform scale before training the ML.

vii) *Data Partitioning:* The preprocessed data from both datasets were partitioned into multiple train-test split ratios such as 80:20, 70:30, and 60:40 to train and evaluate the ML models performance on the testing data.

viii) *Checking for Imbalanced Data:* Oversampling using Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the instances of DM of the target classes in both datasets.

ix) *Feature Selection:* Feature selection was employed using the BORUTA wrapper algorithm to choose the most relevant features towards predicting DM for each of these datasets.

C. Proposed Models

For conducting this research, five ML models encompassing supervised, ensemble, and deep learning were selected after conducting the literature review. These models include the:

i) *Support Vector Machine (SVM)*
ii) *Random Forest (RF).*
iii) *Naïve Bayes (NB).*
iv) *eXtreme Gradient Boosting (XGBoost).*
v) *Deep Neural Network (DNN):* For this model, there were 2 sets of parameters initiated with one set being used to prepare this model for binary classification tasks and the other for multiclass classification tasks. These 2 sets of parameters have been listed down in Tables III and IV.

Table III: Parameters of DNN model for binary classification.

Parameters	Value
Number of Hidden Layers	5
Optimization Algorithm	Adam
Loss Function	Binary Cross Entropy
Number of Epochs	10
Batch Size	64

Table IV: Parameters of DNN model for multiclass classification.

Parameters	Value
Number of Hidden Layers	5
Optimization Algorithm	Adam
Loss Function	Sparse Categorical Cross Entropy
Number of Epochs	10
Batch Size	64

D. Evaluation Metrics Used

The evaluation metrics utilized to evaluate the performance of the ML models are the accuracy, precision, recall and F1-Score.

i) *Accuracy:* The accuracy evaluation metric represents the ratio of correctly classified outcomes to the total classified outcomes.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

ii) *Precision:* Additionally, the precision measure was used to get the accuracy of positive predictions made by the models.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

iii) *Recall:* For computing the models' ability to correctly distinguish positive instances, the recall evaluation metric was utilized.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

iv) *F1-Score:* Finally, the F1-Score was calculated to represent the harmonic mean of the precision and recall measures, providing a balanced measure of performance of each model.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

IV. EXPERIMENTAL RESULTS

In this section, the experimental results of our study are presented, focusing on the prediction of DM using the proposed ML models. We evaluate the performance of these models on two datasets, the PIMA Indian Diabetes and NHANES 1999-2016 datasets using binary classification for both and multiclass classification on the NHANES 1999-2016 dataset only.

A. Oversampling Training Data Using Synthetic Minority Oversampling Technique

i) *PIMA Indian Diabetes Dataset:* The number of instances in the target class before and after applying Synthetic Minority Oversampling Technique (SMOTE) to balance the PIMA Indian Diabetes dataset for binary classification can be seen in Table V. In the training data using a train-test split ratio of 80:20, there were 401 instances of class 0, indicating no DM, and 213 instances of class 1, indicating the presence of DM. After applying SMOTE to this training data, both classes were balanced with 401 instances of each class. Similarly, for the 70:30 and 60:40 train-test split ratios, SMOTE was utilized to successfully balance the training data, ensuring an equal number of instances for both classes.

Table V: Instances of target class in training data set after balancing the PIMA dataset for binary classification using SMOTE.

Training Set Size	**Records before SMOTE**		**Records after SMOTE**	
Target Class Labels	0	1	0	1
80%	401	213	401	401
70%	349	188	349	349
60%	294	166	294	294

ii) *NHANES 1999-2016 Dataset for Binary Classification:* In Table VI, it illustrates the transformation of the NHANES 1999-2016 dataset for binary classification tasks before and after applying SMOTE to balance the training data. In the training data employing a train-test split ratio of 80:20, there were 25,755 instances of class 0, indicating no DM, and 3,908 instances of class 1, indicating the presence of DM. After SMOTE oversampling, the training data of the 80:20, 70:30, and 60:40 train-test split ratios now have an equal number of instances for both classes, with 25,755, 22,527 and 19,304 instances for each train-test ratio respectively, ensuring the fair representation of both classes in the training data.

Table VI: Instances of target class in training data set after balancing the NHANES 1999-2016 dataset for binary classification using SMOTE.

Training Set Size	**Records before SMOTE**		**Records after SMOTE**	
Target Class Labels	0	1	0	1
80%	25755	3908	25755	25755
70%	22527	3428	22527	22527
60%	19304	2943	19304	19304

iii) *NHANES 1999-2016 Dataset for Multiclass Classification:* Table VII shows the impact of SMOTE oversampling on the NHANES 1999-2016 dataset for multiclass classification. The dataset initially contained instances across three classes, with 0 indicating no DM, 1 indicating DM, and 2 indicating borderline DM. SMOTE was applied to balance these classes for the training data of different train-test split ratios. As a result, the number of instances for each class is balanced for all the train-test split ratios of 80:20, 70:30 and 60:40. For example, in the 80:20 train test split ratio, each target class has 25,755 instances after oversampling using SMOTE, ensuring a balanced representation of all classes in the training data. This approach has also been implemented across the other train-test split ratios, allowing for the fair model training of multiclass classification tasks.

TABLE VII: Instances of target class in training data set after balancing the NHANES 1999-2016 dataset for multiclass classification using SMOTE.

Training Set Size	**Records before SMOTE**			**Records after SMOTE**		
Target Class Labels	0	1	2	0	1	2
80%	25755	3366	542	25755	25755	25755
70%	22527	2956	472	22527	22527	22527
60%	19304	2528	415	19304	19304	19304

B. BORUTA Feature Selection

The BORUTA wrapper algorithm, chosen as our feature selection method, enhances model performance by ranking the importance of each feature in the dataset. It employs a Random Forest (RF) classifier to assign scores to features, enabling us to retrain and re-evaluate our machine learning models using only the most relevant features. This process optimizes model accuracy and provides insights into feature relevance.

i) *PIMA Indian Diabetes Dataset:* Table VIII displays BORUTA feature scores for the PIMA Indian Diabetes dataset, focusing on the 60:40 train-test split with the highest overall accuracy. Scores range from 0.00 to 1.00, indicating feature importance for prediction. “Glucose”, “BMI”, “Diabetes Pedigree Function”, and “Age” scored 1.00, highlighting their significance. Conversely, “Insulin” and “BloodPressure” scored lower at 0.00 and 0.33, suggesting limited impact on model performance.

Table VIII: BORUTA feature scores for PIMA Indian Diabetes dataset.

Rank	Features	Feature Score
1	Glucose	1.00
2	BMI	1.00
3	Diabetes Pedigree Function	1.00
4	Age	1.00
5	Pregnancies	0.67
6	Skin Thickness	0.67
7	Blood Pressure	0.33
8	Insulin	0.00

ii) *NHANES 1999-2016 Dataset for Binary Classification:* Table IX presents BORUTA feature scores for NHANES 1999-2016 binary DM classification, focusing on the 80:20 train-test split with the highest accuracy. Scores range from 0.00 to 1.00, indicating feature importance. Features like “Age”, “GGT”, “Red-Cell-Distribution-Width”, and more scored 1.00, highlighting their significance. Conversely, “Moderate Work” and “Gender” had lower scores at 0.08 and 0.00, indicating limited impact on model performance.

Table IX: BORUTA feature scores for NHANES 1999-2016 dataset for binary classification.

Rank	Features	Feature Score
1	Age	1.00
2	GGT	1.00
3	Red-Cell-Distribution-Width	1.00
4	X60-sec-pulse	1.00
5	ALP	1.00
6	Cholesterol	1.00
7	Creatinine	1.00
8	Glucose	1.00
9	Iron	1.00
10	Platelet-count	1.00
11	LDH	1.00
12	Uric Acid	1.00
13	Triglycerides	1.00
14	Total-Cholesterol	1.00
15	HDL	1.00
16	Glycohemoglobin	1.00
17	Hematocrit	1.00
18	Albumin	1.00
19	White-Blood-Cells	1.00
20	Red-Blood-Cells	1.00
21	Systolic	1.00
22	Diastolic	1.00
23	Weight	1.00
24	Height	1.00
25	Body-Mass-Index	1.00
26	Lymphocyte	1.00
27	Blood-Rel-Diabetes	1.00
28	Hemoglobin	1.00
29	Mean-Cell-Hgb-Conc.	1.00
30	Mean-Cell-Vol	1.00
31	AST	0.93
32	Segmented-Neutrophils	0.86
33	Mean-cell-Hemoglobin	0.79
34	Mean-Platelet-Vol	0.71
35	Monocyte	0.64
36	Eosinophils	0.57
37	Vigorous-work	0.50
38	ALT	0.43
39	Phosphorus	0.36
40	Protein	0.29
41	Bilirubin	0.21
42	Basophils	0.16
43	Moderate-work	0.07
44	Gender	0.00

iii) *NHANES 1999-2016 Dataset for Multiclass Classification:* Table X displays BORUTA feature scores for NHANES 1999-2016 multiclass DM classification using an 80:20 train test split with the highest accuracy. Scores between 0.00 and 1.00 signify feature importance. Features like “Age”, “Platelet-count”, “Glycohemoglobin”, and “Total-Cholesterol” scored 1.00, crucial for accurate multiclass DM prediction. Conversely, “Moderate-Work” and “Gender” had lower scores at 0.03 and 0.00, indicating minimal impact on multiclass classification performance.

Table X: BORUTA feature scores for NHANES 1999-2016 dataset for multiclass classification.

Rank	Features	Feature Score
1	Age	1.00
2	Platelet-count	1.00
3	Glycohemoglobin	1.00
4	Total-Cholesterol	1.00
5	Triglycerides	1.00
6	Iron	1.00

7	Glucose	1.00
8	Cholesterol	1.00
9	Hematocrit	1.00
10	Blood-Rel-Diabetes	1.00
11	Weight	1.00
12	Systolic	1.00
13	Body-Mass-Index	1.00
14	Uric Acid	0.97
15	Red-Cell-Distribution-Width	0.94
16	Hemoglobin	0.94
17	Diastolic	0.87
18	HDL	0.84
19	Red-Blood-Cells	0.81
20	Height	0.77
21	ALP	0.74
22	Creatinine	0.71
23	X60-sec-pulse	0.68
24	LDH	0.68
25	Monocyte	0.61
26	Lymphocyte	0.58
27	Segmented-Neutrophils	0.55
28	Eosinophils	0.52
29	Mean-Cell-Hgb-Conc.	0.48
30	ALT	0.45
31	Albumin	0.45
32	White-Blood-Cells	0.39
33	AST	0.35
34	GGT	0.29
35	Mean-Platelet-Vol	0.29
36	Mean-Cell-Vol	0.29
37	Mean-cell-Hemoglobin	0.23
38	Phosphorus	0.19
39	Protein	0.16
40	Bilirubin	0.13
41	Basophils	0.10
42	Vigorous-work	0.06
43	Moderate-work	0.03
44	Gender	0.00

C. Binary Classification of Diabetes using PIMA Indian Diabetes Dataset

The binary classification of DM on the PIMA Indian Diabetes dataset using all the features in the dataset after pre-processing and train-test split ratios of 80:20, 70:30 and 60:40. It was found that the 60:40 train test split ratio achieved the best results using all features. After employing BORUTA feature selection, the best features were obtained, and the models were then re-evaluated using these features and the 60:40 train-test split ratio. Out of all the possible variations, the RF model achieved the highest accuracy of 79% out of all the models after being trained and evaluated using the best features from BORUTA feature selection and a train-test split ratio of 60:40. The results in the form of the calculated evaluation metrics, are presented in Tables XI and XII.

Table XI: Results obtained after evaluating models on testing data after training using all features of PIMA Indian Diabetes dataset and various train-test split ratios (Binary Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 80:20				
SVM	70%	56%	73%	63%
RF	77%	66%	76%	71%
NB	69%	55%	75%	64%
XGBoost	71%	57%	75%	65%
DNN	66%	51%	78%	62%
Train-Test Split of 70:30				
SVM	72%	58%	69%	63%
RF	77%	65%	75%	69%
NB	68%	53%	71%	61%
XGBoost	71%	57%	68%	62%
DNN	69%	54%	71%	62%
Train-Test Split of 60:40				
SVM	74%	59%	70%	64%
RF	78%	64%	75%	69%
NB	69%	53%	71%	61%
XGBoost	71%	55%	66%	60%
DNN	69%	52%	74%	61%

Table XII: results obtained after evaluating models on testing data after training using best features and 60:40 train-test split ratio of PIMA Indian Diabetes dataset (Binary Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 60:40				
SVM	77%	63%	76%	79%
RF	79%	65%	77%	71%
NB	75%	60%	72%	65%
XGBoost	75%	60%	73%	66%
DNN	70%	53%	78%	63%

D. Binary Classification of Diabetes using NHANES 1999-2016 Dataset

Subsequently, the binary classification of DM on the NHANES 1999-2016 dataset using all the features in the dataset after preprocessing and train-test split ratios of 80:20, 70:30 and 60:40. It was found that the 80:20 train-test split ratio achieved the best results using all features. After employing BORUTA feature selection, the best features were obtained, and the models were then re-evaluated using these features and the 80:20 train-test split ratio. Out of all the possible variations, the XGBoost model achieved the highest accuracy of 92% out of all the models after being trained and evaluated using all the features obtained after BORUTA feature selection and a train-test split ratio of 80:20. The results in the form of the computed evaluation metrics, are presented in Tables XIII and XIV.

Table XIII: Results obtained after evaluating models on testing data after training using all features of NHANES 1999-2016 dataset and various train-test split ratios (Binary Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 80:20				
SVM	82%	37%	58%	45%
RF	90%	58%	62%	60%
NB	74%	28%	68%	40%

XGBoost	92%	73%	56%	63%
DNN	78%	34%	74%	46%
Train-Test Split of 70:30				
SVM	82%	36%	57%	44%
RF	89%	58%	60%	59%
NB	74%	29%	69%	41%
XGBoost	91%	71%	56%	62%
DNN	78%	34%	71%	46%
Train-Test Split of 60:40				
SVM	82%	37%	57%	45%
RF	89%	59%	60%	59%
NB	75%	29%	69%	41%
XGBoost	91%	70%	56%	62%
DNN	81%	36%	66%	47%

Table XIV: results obtained after evaluating models on testing data after training using best features of NHANES 1999-2016 dataset and 80:20 train-test split ratio (Binary Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 80:20				
SVM	81%	35%	64%	45%
RF	89%	57%	66%	61%
NB	74%	29%	73%	42%
XGBoost	92%	71%	58%	63%
DNN	77%	32%	75%	45%

E. Binary Classification of Diabetes using NHANES 1999-2016 Dataset

Lastly, multiclass classification of DM on the NHANES 1999-2016 dataset was carried out. It is only implemented on this dataset because it contained more than 2 target classes for DM. It was found that the 80:20 train-test split ratio achieved the best results using all features. After employing BORUTA feature selection, the best features were obtained, and the models were then re-evaluated using these features and the 80:20 train-test split ratio. Out of all the possible variations, the XGBoost model achieved the highest accuracy of 91% out of all the models after being trained and evaluated using all the features obtained after BORUTA feature selection and a train-test split ratio of 80:20. The results in the form of the computed evaluation metrics, are presented in Tables XV and XVI.

Table XV: Results obtained after evaluating models on testing data after training using all features of NHANES 1999-2016 dataset and various train-test split ratios (Multiclass Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 80:20				
SVM	81%	43%	48%	45%
RF	88%	50%	55%	52%
NB	67%	43%	51%	42%
XGBoost	91%	59%	53%	54%
DNN	77%	46%	51%	46%
Train-Test Split of 70:30				
SVM	81%	44%	49%	46%
RF	89%	54%	55%	53%
NB	67%	43%	51%	42%
XGBoost	91%	54%	52%	53%
DNN	69%	45%	54%	44%
Train-Test Split of 60:40				
SVM	81%	44%	48%	45%
RF	88%	50%	53%	51%
NB	67%	43%	52%	42%
XGBoost	91%	56%	52%	53%
DNN	71%	44%	54%	44%

Table XVI: Results obtained after evaluating models on testing data after training using best features of NHANES 1999-2016 dataset and 80:20 train-test split ratio (Multiclass Classification).

Model	Accuracy	Precision	Recall	F1-Score
Train-Test Split of 80:20				
SVM	70%	44%	53%	44%
RF	86%	50%	57%	53%
NB	67%	44%	54%	43%
XGBoost	90%	57%	55%	54%
DNN	70%	44%	55%	44%

F. Comparison with Previous Works

In this section, we compare the accuracy of our models with the previous works of other researchers on the same datasets.

i) *Binary Classification of PIMA Indian Diabetes Dataset:* Table XVII presents the comparison of model accuracies on the PIMA Indian Diabetes dataset with previous works. The proposed method achieved competitive accuracy levels when compared to the previous works but did not achieve a best accuracy indicating that there are better methods that can be implemented.

Table XVII: Comparison of best model accuracy achieved on Binary Classification of PIMA Indian Diabetes dataset with previous works done.

Previous Work	Models Employed	Best Accuracy Achieved
Proposed Method	RF	79%
Charitha *et al.* [14]	SVM, XGBoost	81.1%
Sarwar *et al.* [9]	SVM	77%
Sonar *et al.* [13]	SVM, Artificial NN	82%

ii) *Binary Classification of NHANES 1999-2016 Dataset:* Table XVIII presents the comparison of model accuracies on the NHANES 1999-2016 dataset with previous work done. Our proposed method outperformed some of the best performing models used in previous works, showcasing the robustness and success of our methodology in binary DM classification on the NHANES 1999-2016 dataset.

Table XVIII: Comparison of best model accuracy of Binary Classification of NHANES 1999-2016 dataset with previous works done.

Previous Work	Models Employed	Best Accuracy Achieved
Proposed Method	XGBoost	92%
Hasan *et al.* [4]	RF	90%

iii) *Multiclass Classification of NHANES 1999-2016 Dataset:* Unfortunately, previous works on the multiclass classification of DM on the

NHANES 1999-2016 dataset were not found, hence no comparison of results could be made.

V. CONCLUSION

In conclusion, this project successfully developed and evaluated multiple ML models for the prediction of DM using medical data. The RF model demonstrated the best accuracy of 79% in binary DM classification on the PIMA Indian Diabetes dataset, employing a 60:40 train-test split ratio and the best from BORUTA feature selection. Meanwhile, the XGBoost model achieved remarkable accuracy scores of 92% and 91% in binary and multiclass DM classification, respectively, on the NHANES 1999-2016 dataset, utilizing an 80:20 train-test split ration using all features in the preprocessed data.

These results align with the project objectives, showcasing the efficacy of RF and XGBoost in DM prediction. Challenges in the form of model selection were addressed through extensive literature review, and time constraints were managed through careful project planning. Future work might include hyperparameter tuning, cross-validation, and testing on additional real-world datasets to further enhance model performance and robustness in real-world scenarios.

REFERENCES

[1] Diabetes, *World Health Organisation (WHO)*, https://www.who.int/news-room/fact-sheets/detail/diabetes. [Retrieved 13 April 2023]

[2] S. U. Jeong, D. G. Kang, D. H. Lee, K. W. Lee, D. M. Lim, B. J. Kim, K. Y. Park, H. J. Chin, G. P. Koh, "Clinical Characteristics of Type 2 Diabetes Patients According to Family History of Diabetes," *Korean Diabetes J.*, 34, pp. 222-228, 2010.

[3] R. Deo, S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," in *IEEE Healthcare Innov. and Point of Care Technol.*, Bethesda, USA, pp. 147-150, 2019.

[4] K.A. Hasan, M. A. M. Hasan, "Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance," in *23rd Int. Conf. Comp. and Inform. Technol.*, Bangladesh, pp. 1-6, 2020.

[5] M. F. Faruque, Asaduzzaman, I. H. Sarker, (2019). "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," in *2019 Int. Conf. Electr., Comp. and Commun. Eng.*, pp. 1-4, 2019.

[6] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, W. Medhat, "Diabetes Prediction Using Machine Learning: A Comparative Study," in *3rd Novel Intellig. and Leading Emerging Sci. Conf.*, Giza, Egypt, pp. 279-282, 2021.

[7] A. C. Lyngdoh, N. A. Choudhury, S. Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithms," in *2020 IEEE-EMBS Conf. Biomedical Eng. and Sci.*, Langkawi Island, Malaysia, pp. 517-521, 2020.

[8] P. K. Saha, N. S. Patwary, I. Ahmed, "A Widespread Study of Diabetes Prediction Using Several Machine Learning Techniques," in *22nd Int. Conf. Comp. and Inform. Technol.*, Dhaka, Bangladesh, pp. 1-5, 2019.

[9] M. A. Sarwar, N. Kamal, W. Hamid, M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in *24th Int. Conf. Automat. and Comput.*, Newcastle Upon Tyne, UK, pp. 1-6, 2018.

[10] G. Tripathi, R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in *8th Int. Conf. Reliability, Infocom Technol. and Optimiz. (Trends and Future Directions)*, Noida, India, pp. 1009-1014, 2020.

[11] P. S. Kohli, S. Arora, "Application of Machine Learning in Disease Prediction," in *4th Int. Conf. Comput. Commun. and Automat.*, Greater Noida, India, pp. 1-4, 2018.

[12] M. Rahman, L. Islam, "Diabetes Recognition in Pregnant Women by Extracting Features Using PCA and Data Mining Algorithms," in *IEEE Pune Sect. Int. Conf.*, Pune, India, pp. 1-6, 2019.

[13] P. Sonar, K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," in *3rd Int. Conf. Comput. Methodol. and Commun.*, Erode, India, pp. 367-371, 2019.

[14] C. Charitha, A. Devi Chaitrasree, P. C. Varma, C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," in *Int. Conf. Comp. Commun. and Inform.*, Coimbatore, India, pp. 1-5, 2022.