Prediction of Micro Vascular and Macro Vascular Complications in Type-2 Diabetic Patients using Machine Learning Techniques

Bandi Vamsi[1]
Department of Computer Science
Artificial Intelligence and Data Science
Madanapalle Institute of Technology and Science
Andhra Pradesh - 517325, Madanapalle, India

Ali Al Bataineh[2]
Department of Electrical
and Computer Engineering
Norwich University
VT 05663, United States

Bhanu Prakash Doppala[3]
Lead Instructor, Data Analytics
Academy Xi
Sydney NSW 2000, Australia

***Abstract*—A collection of metabolic conditions known as diabetes mellitus (DM) is defined by hyperglycemia brought on by deficiencies in insulin secretion, action, or both. In terms of mortality rate, type-2 diabetes is 20 times higher when compared with type-1. Based on the earlier research, there is still scope to identify different risk levels of type-2 diabetes complications. To achieve this, we have proposed a T2DC machine learning-based prediction system using a decision tree as a base estimator with random forest to identify the severity of T2-DM complications at an early stage. Our proposed model achieved accuracies of 95.43%, 94.62%, 96.25%, 97.55%, and 97.83% for Nephropathy, Neuropathy, Retinopathy, Cardio Vascular and Peripheral Vascular complications in T2-DM patients. The proposed model has the potential to improve clinical outcomes by promoting the delivery of early and personalized care to T2-DM patients.**

***Keywords*—Diabetes mellitus; micro vascular; macro vascular; machine learning; type-2 complications**

I. Introduction

One of the most widespread health problems affecting all age groups is diabetes mellitus (DM). According to WHO (World Health Organization) statistics, nearly 180 million people worldwide have type 2 diabetes mellitus (T2-DM), with 95 percent having DM in this structure [1]. The number of people having this T2-DM is estimated to rise drastically by 2030. As per the highest cases recorded in the world, India ranked in 2nd place with 60 million DM records and is estimated to rise by 109 million people with DM by 2035 [2]. It is a condition that is identified when the pancreas fails to produce insulin in the required amount needed for the body, or due to damage to tissues and cells in the human body. T2-DM is a condition that is strongly connected to both micro vascular (MIV) and macro vascular (MAV) problems, which include nephropathy, retinopathy, neuropathy (MIV), and peripheral vascular disease (MAV), contributing to the effect on internal organs and blood vessel-related complications [3]. DM can be identified in three different forms, namely: type 1 diabetes mellitus (T1-DM) affects the pancreas by producing insulin in a lower amount than needed by the entire human body [4]. To keep the body's insulin level at the right level, it needs supplements from outside the body. T2-DM is a condition characterized by a significant increase in blood sugar levels. In this structure of DM, the insulin levels are disturbed, and the body fails to utilize and produce [5]. It is a type of hormone developed in the pancreas that aids in maintaining sugar levels in the blood. In particular, this hormone maintains the amount of glucose that flows through the cells. In general, after the consumption of food, blood sugar levels in the blood are identified in the high range [6]. The extra glucose in the blood is transferred into the cells when the pancreas secretes insulin, which diminishes the quantity of glucose in the bloodstream [7]. When this abnormality is not identified at an early stage using proper diagnosis, this T2-DM can lead to severe chronic health disorders. Chronic hyperglycemia, a side effect of diabetes, can damage, weaken, or kill many organs over time, especially the kidneys, eyes, heart, nerves, and blood vessels [8].

A. MIV and MAV Complications

Based on the research on DM, it is evident that the following are the damaging factors for human health:

- latent loss of eyesight with Retinopathy (RET).
- fluid accumulation causes Nephropathy (NEP), which in turn leads to hyperglycemia.
- Neuropathy (NEU) is a condition that affects the supply of blood by causing late healing in the functioning of nerves, shortening the sensation in the feet, and ulcers.
- the obstruction and shortened flow of oxygenated blood to the bladder and kidneys are caused due to cardio vascular disease (CVD).
- the internal layers of large and small arteries are the complications caused due to peripheral vascular diseases (PVD).

B. Role of Machine Learning in Diabetes Detection

Machine Learning (ML) models [9] have been developed in most medical implementations as an encouraging tool to help in taking spontaneous conclusions related to various infections, along with DM, which produces favorable results. With ML algorithms, vast amounts of data are processed by minimizing the effort [10]. This data is used to train models, which then generate the most appropriate results associated

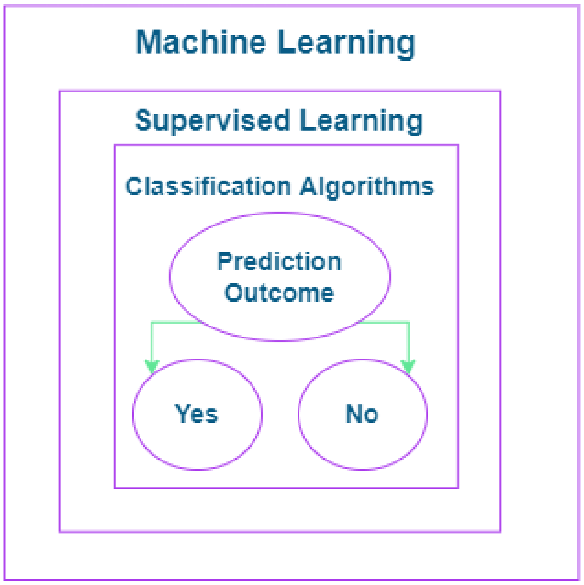


Fig. 1. Machine Learning Classification.

with the input data [11]. It is possible to train the models using any characteristics that are feasible and medically necessary. These parameters differ in accordance with the wide range of symptoms. Some of the learning models used for classification purposes are decision trees, support vector machines, and regression analysis [12]. The approaches are often embedded in statistical analysis to extract useful information from the core data. A combination of these approaches can be used to build a predictive model that identifies the risk complications of T2-DM based on the deciding risk prediction calculator. A scenario of disease prediction with machine learning models is depicted in Fig. 1.

This research study focuses on identifying the risk levels of MIV and MAV complications in T2-DM using different ML algorithms. Accurate prediction of MIV and MAV complications could aid in more targeted measures that would prevent or slow their progression.

The rest of this paper has the following structure. Section II discusses related research and its limitations. Section III presents the proposed methodology. Section IV discusses the experimental results, and Section V concludes the study with potential future work.

II. Related Work

Allen A. et al. [13] trained two ML algorithms to predict DKD severity stages. To assess performance, they compared them to the Centers for Disease Control and Prevention (CDC) risk score. The algorithms were validated using both a hold-out test set and an external dataset obtained from different facilities. In both the hold-out and external datasets, their proposed algorithms outperformed the CDC risk score, achieving an area under the receiver operating characteristic curve (AUROC) of 0.75 on the hold-out set for the prediction of any-stage DKD and an AUROC of over 0.82 for more severe endpoints, compared to the CDC risk score, which had an AUROC of less than 0.70 on all test sets and endpoints. Lu H. et al. [14] suggested that the perception of acute infection changes and predicting people inculcating the threat of treatment-resistant infection are mainly considered. For chronic disease identification, an ensemble of original patient-channel and ML techniques is proposed. This proposed method is used in networks with health scenarios. T2-DM is identified in a subset of patients for this purpose. This method identified the factors in identifying the acute infection threat using eight ML techniques. The significant observations show that the advanced structure with ML classifiers achieved an AUC ranging from 0.79 to 0.91. Rashid, M. et al. [15] identified a root cause of death among T2-DM patients due to micro vascular problems. Their study aims to examine the use of the entire ML procedure in identifying issues using people's medical, clinical, and statistical examinations. The records of 96 people from Bangladesh were examined with T2-DM. They are examined through a chi-squared examination to demographically represent the major key points in identifying the micro vascular problems like CAN, DPN, and RET. Various ML models like LR, RF, and SVM were used for the examination of micro vascular problems. The exact outcomes are determined through the random forest through hypertension, gender, micro albuminuria, and smoking habit. The authors showed ML represents accurate results in identifying micro vascular problems in T2-DM patients. Based on their records, which aid in controlling these people by later micro vascular problems that lead to early death. Deberneh, H. et al. [16] considered factors like FPG, triglycerides, HbA1c, gamma-GTP, BMI, family history, physical activity, smoking, drinking, gender, age, and uric acid in their study. Then the engaged LR, RF, SVM, XGBoost and ensemble ML procedures relied on these attributes to identify the result as normal, diabetic, or pre-diabetic. Depending on the hypothetical outcomes, the execution of the identified method strives to preferably better in predicting the circumstance of type 2 diabetes. This method also helps doctors and patients with required forecasting data on the probability of occurring type 2 diabetes. Fazakis et al. [17] proposed a worker-centric, IoT-enabled, unobtrusive health, well-being, and functional ability monitoring framework with AI tools for the early detection of T2-DM. Their diabetes risk prediction system used several ML models to apply, evaluate, and incorporate KDD components. The ensemble WeightedVotingLRRFs ML model's AUC of 0.884 improves diabetes prediction. Jian, Y. et al. [18] proposed multiple ML algorithms to predict and classify eight diabetes complications. Metabolic syndrome, dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity, and retinopathy are among the complications. The authors used a dataset with 884 cases and 79 features. The models' performance was evaluated using accuracy and F1-score metrics, which reached a maximum of 97.8% and 97.7%, respectively. Neha Prerna et al. [19] proposed a research study that uses different ML algorithms to predict the risk of type 2 diabetes among individuals based on their lifestyle and family history. The experiment was carried out with 952 instances collected via an online and offline questionnaire, which included 18 questions about health, lifestyle, and family background. The proposed ML algorithms were also tested on the "Pima Indian Diabetes database." Their experimental results revealed that the random forest classifier performed the best in terms of accuracy. Jung, L. C et al. [20] described a method for identifying subtle effects of genetic variants using whole genome sequencing data and improving the prediction accuracy of T2-DM at the population level. The method entailed first performing sparse principal component analysis to genotype data to obtain orthogonal features, then creating a new classifier with single nucleotide polymorphism (SNP)-specific regularization parameters to reduce the false positive rate of feature selection, and finally verifying feature relevance with penalized logistic regression. The researchers used a dataset containing 625597 SNPs and 23 environmental variables from 3326 people. The method identified 271 genetic variants with minor effects on T2DM prediction. It is also more than 15 times faster than random forest and extreme gradient boosting (XGBoost) classifiers. Hasan, M.K. et al. [21] proposed a method for DM identification for the metadata, involving 768 women, 268 of who had high blood sugar levels and 500 of whom were normal. In their study, initialization is critical for maintaining cutting-edge results. This includes edge elimination, replacement with the average for lost numbers, data stabilization, factor alternative, and five-fold validation. K-Nearest Neighbor, Decision Trees, Random Forest, AdaBoost, and other Different ML models were used. The authors also proposed a weighted ensembling of different ML models. AUC was chosen as the performance metric. The experiments demonstrated that the ensembling classifier outperforms all others, with sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC values of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively. Islam, M.S. et al. [22] proposed a study to predict the Hemoglobin Alc (HbAlc) levels in advance using ML methods in order to enable early diagnosis and prevent diabetes complications. The fractional derivative, glucose variability, time in range, and wavelet decomposition methods were used to extract features from continuous glucose monitoring (CGM) data. The CGM data from the Diabetes Research in Children Network (DirecNet) was utilized. According to the results, the ensembling of the random forest and extreme gradient boosting algorithms, combined with the feature fusion, produced the best performance with a low mean absolute error (MAE) of 3.39 mmol/mol and a high coefficient of determination (R-squared) score of 0.81. Kopitar, L. et al. [23] compared ML models for the prediction of T2-DM by using various regression techniques on undiagnosed patient data. Fasting plasma samples are measured over a six-month period using 100 computational iterations. These iterations were examined using various data subsets. According to the study analysis, the linear regression model had the lowest average RMSE of 0.838, followed by random forest with 0.842 and Xgboost with 0.881. Dagliati et al. [24] proposed a data mining pipeline to derive a set of predictive models of T2DM complications from nearly 1000 patients' electronic health record data. Clinical center profiling, predictive model targeting, predictive model construction, and model validation are all part of the pipeline. The logistic regression-based method with stepwise feature selection was used to predict the onset of retinopathy, neuropathy, or nephropathy at three different time intervals: three, five, and seven years after the initial visit to the Hospital Center for Diabetes. Gender, age, time since diagnosis, BMI, hypertension, HbA1c, and smoking habit were all factors considered in the study. The final models, customized for the complications, had an accuracy of up to 0.838. For each complication and time scenario, different attributes were chosen. This led to specialized models that are easy to use in clinical practice. Wei, S. et al. [25] used a variety of machine learning techniques, i.e., Neural Networks, Support Vector Machines, and Decision Trees to detect diabetes. The best accuracy acquired was 77.86% through the 10-fold cross-validation approach. Fan Yuting et al. [26] proposed an effective ML model for identifying the problem of blood sugar levels in non-adherent T2-DM. The authors looked at people who had not had glycosylated hemoglobin in the previous month to identify the risk in blood sugar levels. Seven different ML procedures are utilized to implement eighteen identification methods. Identification achievement is majorly analyzed through the AUC of the examining group. Based on 800 patients' data, 20.6% could meet the insertion range, of which 78.2% had poor glycemic guide. The greater AUC of the analysis set for nephropathy, peripheral neuropathy, angiopathy, retinopathy, and glycosylated hemoglobin is determined as 0.902 ± 0.040, 0.859 ± 0.050, 0.889 ± 0.059, 0.832 ± 0.086, and 0.825 ± 0.092 accordingly. Both the ML models and univariate testing attained an equal outcome. Fiarni et al. [27] proposed Naive Bayes and C4.5 classification approaches, as well as k-means clustering, for identifying the risk complications of T2-DM patients. The authors analyzed each technique's reliability and identified the associated elements and sub-features as clinical contributing factors. As a result, the most major risk factor for retinopathy is a female patient who is now experiencing a hypertensive problem. In terms of nephrotic syndrome, the most major risk factor is a history of diabetes lasting over four years. Furthermore, it was more prevalent in female patients with a BMI over 25. There is no clear association between the duration of diabetes and certain complications. The overall accuracy of the suggested model is 68%, which means that it could be used as an alternative way to find diabetes complications early. Sudharsan, B. et al. [28] proposed a reducing phenomenon of blood sugars in people having T2-DM. The authors analyzed the risk of complications through different data sets. The quantity of self-regulation of blood sugar levels required by the method is nearly ten per week. The vulnerability of the method for identifying blood sugar levels in the coming 24 hours is 92%, and the selectivity is 70%. In their work, the identification group was four hours of blood sugar, and the selectivity advanced to 90%. The advanced ML methods can identify blood sugar levels with an accurate level of vulnerability and selectivity.

After taking earlier research work into consideration, we observed that the majority of the work concentrated on identifying the T2-DM with limited complications. To overcome this limitation and to identify the severity levels, in this proposed work, we concentrated on predicting the risk levels at an early stage, falling under low, medium, and high categories. These levels are discriminated against among the T2-DM patients with MIV and MAV complications such as Nephropathy, Neuropathy, Retinopathy, cardio vascular and Peripheral vascular.

III. Methodology

This section goes into more detail about the statistical analysis of the dataset used for this study, the T2-DM patient network, data preprocessing approaches, ML methods, and the proposed model.

A. Statistical Analysis of the Dataset

For this study, 3068 records are considered, particularly with subjects between 30 and 80 years of age, of which 1565 are males, and 1503 are females. The mean, standard deviation, error rate, and P-value of primary attributes that are available in the dataset are displayed in Table I. These attributes are mainly considered in identifying the risk levels of MIV and MAV complications in T2-DM patients collected from various multi-specialty hospitals in India [29].

The distribution of samples depends on the parameters in a dataset and can be represented using frequency tables. These are useful in making decisions that appear more or less within the dataset. Every parameter and its selection range can be easily identified. From this, basic charts like histograms and bar charts can be generated for data visualization. The representation of some primary attributes in the dataset is depicted using a histogram, shown in Fig. 2.

B. T2-DM Complications Network

In this work, the graph theory concepts are used to construct the T2-DM complications network. This network represents the common complications among different patients. For this, four different patients with MIV and MAV complications are considered from the dataset. The relation among patients and various complications are determined by Eq. (1).

$$G = (V, E) \tag{1}$$

where ‘V’ represents the nodes that indicate the patient’s complications. The common complications for these patients are connected through ‘edges’ (E). Hence, these sets of nodes and edges are combined together to form a complication network (G).

Fig. 3. represents the complication network among T2-DM patients to identify the relationship between patients and their common health complications. The left-hand side of this figure shows three male and one female patients’ data along with their disease complications, respectively. The middle part represents the edge relation between patients connected through nodes. For instance, patients p1, p2, and p4 all have T2-DM in common. Due to this, ”neuropathy” is a common complication. Also, patients p2 and p4 show a common complication of ”nephropathy”. Hence, by considering this scenario, the patient’s p1, p2, and p4 are connected through nodes, and their common health comorbidities are interlinked with edges, thereby forming a subgraph to construct a complications network. By constructing this network, the classification between T2-DM and non-T2-DM becomes easier through ML models.

C. Data Preprocessing

When developing ML models, data preprocessing is the first step in the process. Real-world data is tainted and contaminated by inconsistencies, noise, incomplete information, missing values, inaccurate (containing errors or outliers), and lack of specific attribute values [31]. This is where data preprocessing comes into play; it helps to clean, format, and organize raw data, preparing it to build ML models. Simply put, data preprocessing helps improve data quality and promotes the extraction of meaningful insights from data to train more accurate prediction models. The data preprocessing procedure in ML includes the following steps [32].

- **Missing Values:** This step involves identifying and appropriately handling missing values; failing to do so could lead to inaccurate and erroneous conclusions and inferences drawn from the data. There are two methods for dealing with missing data: (a) deleting a specific row, in which we remove a particular row that contains a null value for a feature or a particular column where more than 70% of the values are missing; and (b) calculating the mean, median, or mode of a specific feature, column, or row that contains a missing value and replacing the outcome with the missing value. This method is helpful for features with numeric data, such as salary, year, and so on, and it can add variance to the dataset while efficiently negating any data loss. As a result, it produces better results than the first method.
- **Removing Duplicates:** This step includes deleting duplicate entries. During model training, an entry that appears more than once is given disproportionate weight. Where identical entries are not all in the same set, duplicate entries can ruin the split between training, validation, and test sets. This can lead to biased estimates of how well the model will perform, which can cause the model to underperform in production.
- **Removing irrelevant data:** This step involves removing irrelevant entries from the dataset. Data often comes from a variety of sources, and a given set of data is likely to have entries that don’t belong.
- **Detecting Outliers:** This step entails detecting outliers by exploring the ranges and possibilities for categorical and numerical data entries. For instance, a negative price for a vehicle is an outlier. Outlier detection or anomaly detection algorithms, such as Isolation Fores or KNN, can also be used to detect and remove outliers automatically.
- **Categorical Data Encoding:** This step includes transforming categorical data (i.e., a patient’s gender) into numerical values. ML models are built on mathematical equations that can only work with numbers. As a result, the categorical values of the features must be converted into numerical values, which can then be fed into ML models to learn from and improve performance.
- **Feature Selection:** This step always plays an important role in machine learning, where we will have several features in the dataset and have to select the best ones when building a model. The inclusion of irrelevant features reduces the model’s generalization capability and may reduce a classifier’s overall accuracy. In addition, the model’s overall complexity grows as more features are added. Feature selection methods in machine learning can be broadly classified as Wrapper, Embedded, and Filter. Wrapper methods use a greedy search approach, evaluating all possible feature combinations against the evaluation criterion. Embedded methods are iterative in the sense that they handle each iteration of the model training process and extract those features which contribute the most to the training for that iteration. Filter methods pick

TABLE I. Summary of Primary Attributes of the Dataset [29]

Parameter	Description	Range	Population Size (n)	Mean	Standard Deviation	Standard Error Mean	*P*-value
Age	Life span of patient	30 to 49	1222	54.92	14.494	0.262	0.011
		50 to 59	600				
		60 to 69	598				
		70 to 80	648				
Sex	Identity of patient	Male	1565	0.51	0.500	0.009	0.048
		Female	1503				
BMI	Determines the level of fat	18.5	155	28.823	6.3138	0.1140	0.025
		18.5 to 24.9	779				
		25.0 to 29.9	736				
		30.0	1398				
SBP	Pressure in arteries when heart beats	120	462	139.32	20.008	0.361	0.015
		120 to 139	891				
		140	1715				
DBP	Pressure in arteries when heart rest among the beats	80	822	84.67	11.348	0.205	0.037
		80 to 89	755				
		90	1491				
HbA1C	Blood pressure attached to hemoglobin	5.7%	127	11.1308	2.671	0.4823	0.016
		=5.7% to 6.4%	617				
		=6.5%	2324				
FBS	Blood sugar level after fasting	=100 mg/dL	40	230.19	76.879	1.388	0.006
		100 to =125 mg/dL	288				
		=126 mg/dL	2740				
PPBS	Determine type of sugar	180 mg/dL	447	334.67	123.539	2.230	0.001
		=180 to 250 mg/dL	492				
		=250 mg/dL	2129				
DIA_LIFE	Span of diabetes in months	40	1350	43.14	12.367	0.223	0.038
		=40 to 60	1388				
		=60	330				
Smoking	-	No	847	1.44	1.129	0.020	0.017
		Ex-smoker	752				
		Occasionally	740				
		Current	729				
Medical Usage	Medicine usage	No	1498	0.51	1.141	0.009	0.019
		Yes	1570				
Medical adherence	Medicine usage pertained to time	Low	1083	0.97	0.822	0.015	0.001
		Medium	994				
		High	991				

TABLE II. Performance of Micro Vascular and Macro Vascular Modules

Module	Model	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Nephropathy (NEP)	RF with base DT	95.43%	96.57%	94.0%	96.81%	95.27%
	RF with base LR	94.78%	97.80%	91.13%	98.12%	94.35%
	RF with base AB	92.91%	93.85%	90.73%	94.81%	92.26%
Neuropathy (NEU)	RF with base DT	94.62%	96.84%	92.0%	97.13%	94.35%
	RF with base LR	93.05%%	94.57%	91.26%	94.81%	92.88%
	RF with base AB	91.12%	93.52%	88.54%	93.75%	90.96%
Retinopathy (RET)	RF with base DT	96.25%	96.78%	95.85%	96.66%	96.32%
	RF with base LR	93.37%	94.83%	92.25%	94.58%	93.53%
	RF with base AB	90.78%	92.65%	88.94%	92.71%	90.76%
Cardio Vascular (CVD)	RF with base DT	97.55%	96.0%	98.96%	96.28%	97.46%
	RF with base LR	93.91%	93.54%	93.95%	93.88%	93.75%
	RF with base AB	91.04%	92.91%	88.63%	93.39%	90.72%
Peripheral Vascular (PVD)	RF with base DT	97.83%	98.0%	86.88%	98.0%	92.11%
	RF with base LR	97.56%	97.9%	71.42%	98.0%	83.33%
	RF with base AB	97.67%	98.0%	78.94%	98.0%	88.23%

up the intrinsic properties of the features measured using univariate statistics. Information Gain and the Chi-square Test are two of the Filter methods. In this research, we have used the Chi-square Test to select the features related to T2-DM. The Chi-square between each feature and the target is calculated, and the number of features with the best Chi-square scores is selected. The Formula for Chi-square is given in Eq. (2), where c is the degrees of freedom, O is the observed value(s), and E is the expected value(s).

$$\chi_c^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (2)$$

- **Data Split:** This step includes splitting the dataset for the ML model into two or more separate sets. Typically, with a two-part split, one part (training dataset) is used to train the ML model, and the other (test dataset) is used to evaluate or test the model. The testing data set is used following training. Usually, the dataset is split into an 80:20 ratio or 70:30 ratio. This

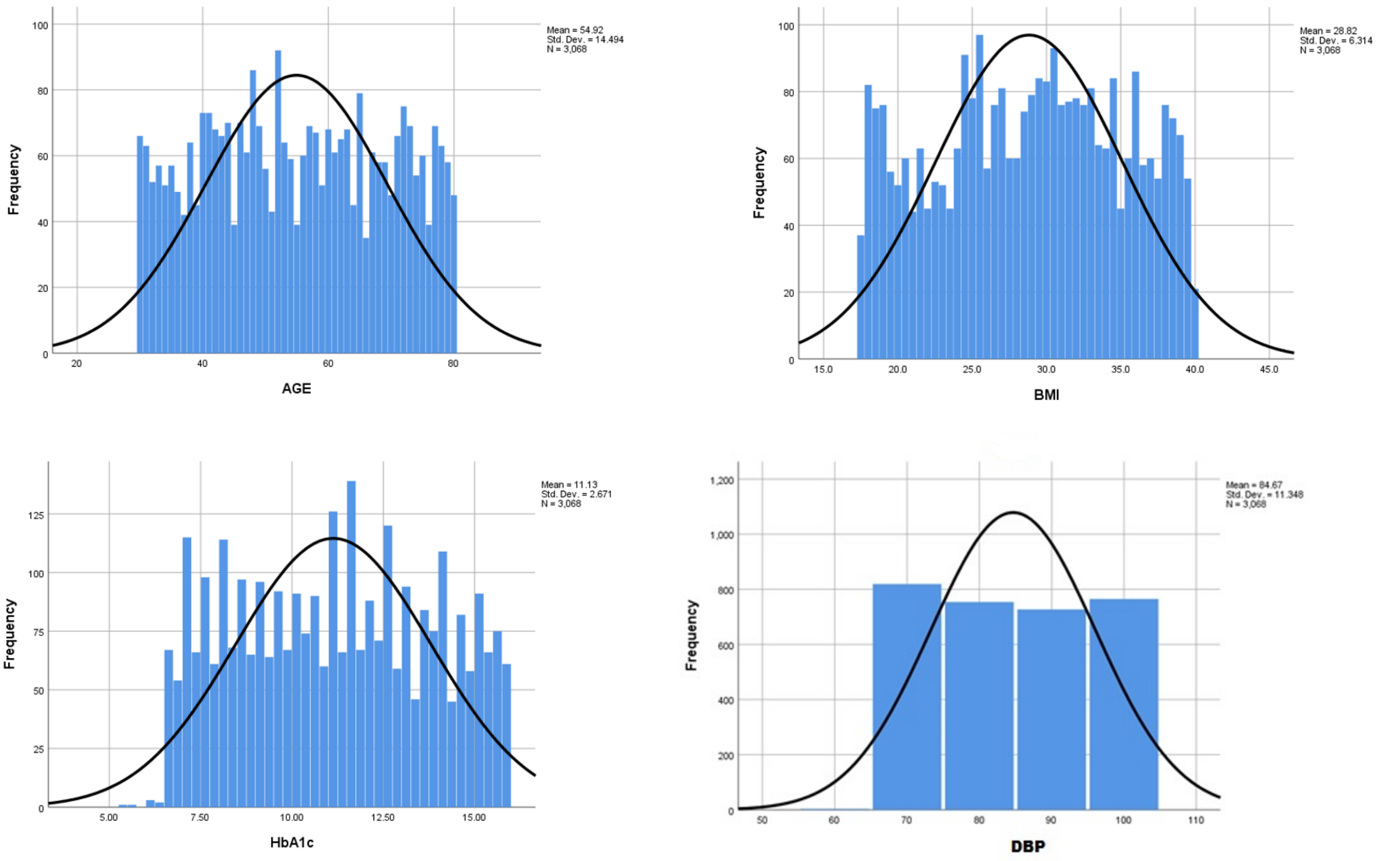


Fig. 2. A Histogram of Some Primary Features in the Dataset.

TABLE III. EXISTING COMPLICATIONS

Patient	Existing Complications				
	NEP	NEU	RET	CVD	PVD
A	0	1	0	0	0
B	1	0	1	0	0
C	0	0	0	0	0
D	1	0	1	0	0
E	1	1	0	0	0

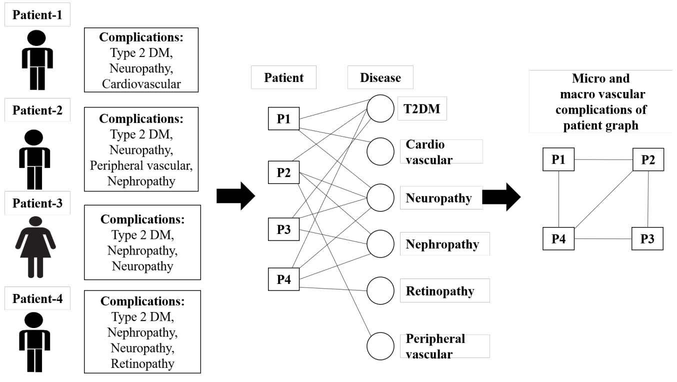

Patient-1
Complications: Type 2 DM, Neuropathy, Cardiovascular
Patient-2
Complications: Type 2 DM, Neuropathy, Peripheral vascular, Nephropathy
Patient-3
Complications: Type 2 DM, Nephropathy, Neuropathy
Patient-4
Complications: Type 2 DM, Nephropathy, Neuropathy, Retinopathy
Patient
Disease
P1
P2
P3
P4
T2DM
Cardio vascular
Neuropathy
Nephropathy
Retinopathy
Peripheral vascular
Micro and macro vascular complications of patient graph
P1
P2
P4
P3

Fig. 3. MIV and MAV Complication Network.

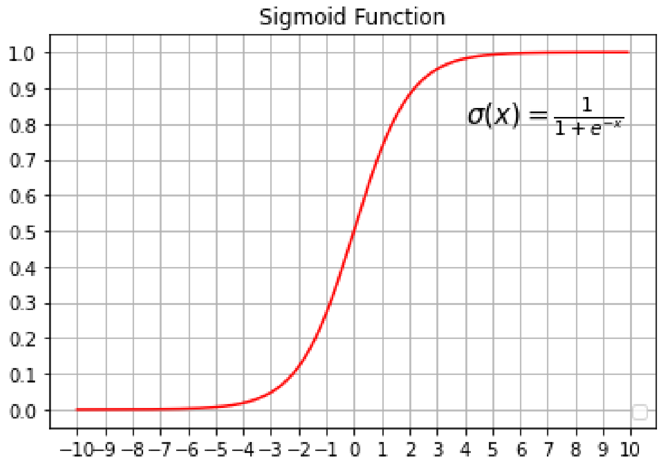
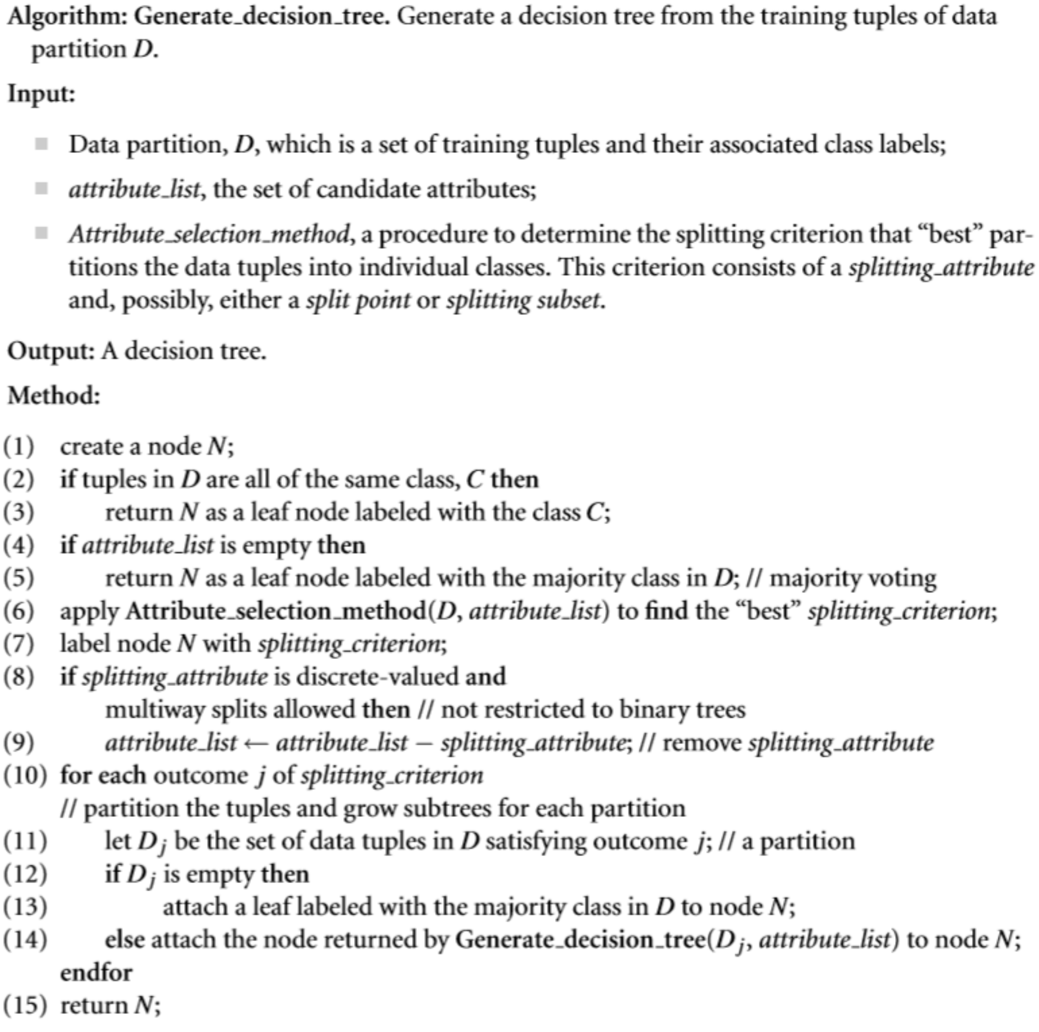


Fig. 4. The Sigmoid Function.

means that 80% or 70% of the data is used for training the model, while the remaining 20% or 30%, is used



Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

```
create a node $N$;
if tuples in $D$ are all of the same class, $C$ then
    return $N$ as a leaf node labeled with the class $C$;
if attribute_list is empty then
    return $N$ as a leaf node labeled with the majority class in $D$; // majority voting
apply Attribute_selection_method($D$, attribute_list) to find the "best" splitting_criterion;
label node $N$ with splitting_criterion;
if splitting_attribute is discrete-valued and
    multiway splits allowed then // not restricted to binary trees
    attribute_list $\leftarrow$ attribute_list $-$ splitting_attribute; // remove splitting_attribute
for each outcome $j$ of splitting_criterion
// partition the tuples and grow subtrees for each partition
    let $D_j$ be the set of data tuples in $D$ satisfying outcome $j$; // a partition
    if $D_j$ is empty then
        attach a leaf labeled with the majority class in $D$ to node $N$;
    else attach the node returned by Generate_decision_tree($D_j$, attribute_list) to node $N$;
endfor
return $N$;
```

Fig. 5. Basic DT Algorithm [30].

TABLE IV. PREDICTED COMPLICATIONS

Patient	Proposed model risk predictions				
	NEP	NEU	RET	CVD	PVD
A	Low	Medium	No	No	No
B	Medium	No	Medium	Low	No
C	Low	No	No	No	No
D	Medium	No	Medium	Low	No
E	Medium	Medium	No	Low	No

for testing. Most often, data is separated into three or more sets. With three sets, the additional set is the validation set, which is used to modify the parameters of the learning process.

- **Data Scaling:** This crucial step concludes the data preprocessing phase in ML. It is a technique for converting all independent features of a dataset to the same scale. This allows for faster learning convergence and more uniform influence across all weights. Normalization and Standardization are two commonly used methods for feature scaling.
    - Normalization: This technique is known as Min-Max scaling, in which all independent feature values are changed between 0 and 1, as defined below:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

where X_{min} and X_{max} are the minimum and the maximum values of the feature, respectively.

    - Standardization: In this technique, the independent feature values are standardized by removing the mean and scaling to unit variance. The standard score of a sample X is calculated as:

$$X' = \frac{X - \mu}{\sigma} \tag{4}$$

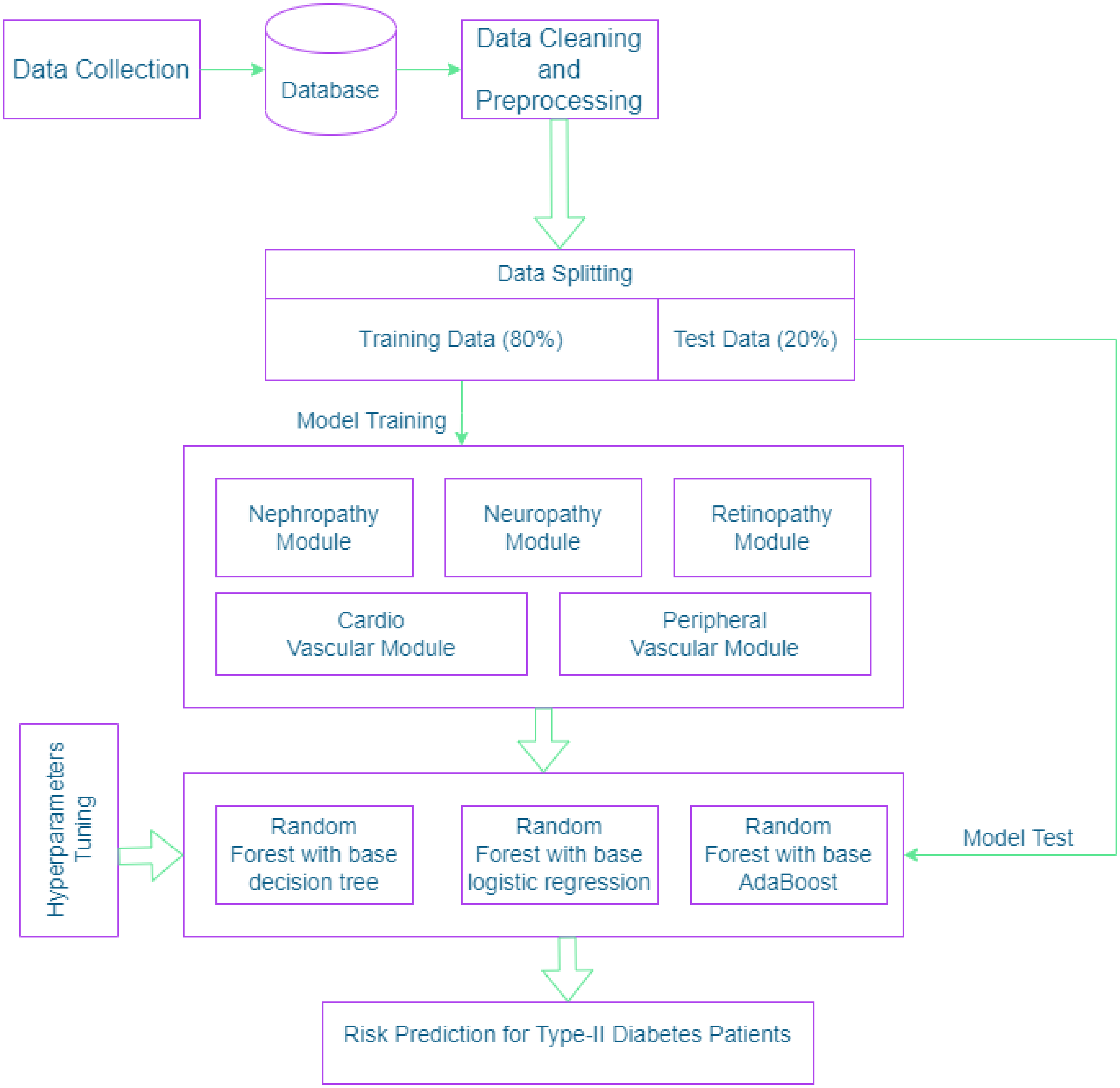

Data Collection
Database
Data Cleaning and Preprocessing
Data Splitting
Training Data (80%)
Test Data (20%)
Model Training
Nephropathy Module
Neuropathy Module
Retinopathy Module
Cardio Vascular Module
Peripheral Vascular Module
Hyperparameters Tuning
Random Forest with base decision tree
Random Forest with base logistic regression
Random Forest with base AdaBoost
Model Test
Risk Prediction for Type-II Diabetes Patients

Fig. 6. Proposed Framework.

where μ is the mean of the feature values, and σ is the standard deviation of the feature values.

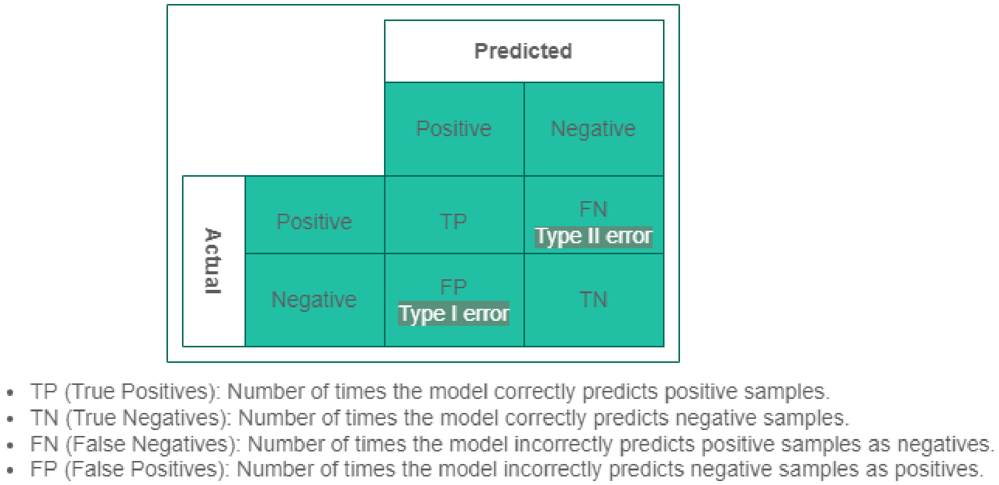
The data preprocessing procedure can vary slightly according to each dataset, but many of the aforementioned steps are applicable to all situations.

D. ML Models

1) Logistic Regression: Logistic Regression (LR) is a powerful ML algorithm commonly used to solve binary classification problems. It is called after the core function of the method, the logistic function. The logistic function (a.k.a. sigmoid function) has an S-shaped curve that can map any real-valued number to a value between 0 and 1 [33]. The sigmoid function, usually denoted by $\sigma(x)$ is defined as follows:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{5}$$

where e is Euler's number and x is the actual numerical value to be transformed. Fig. 4 shows a plot of the numbers between -10 and 10 transformed into the range 0 and 1 using the sigmoid function. LR, like linear regression, uses an equation as its representation. To predict an output value (y),



		Predicted	
		Positive	Negative
Actual	Positive	TP	FN Type II error
	Negative	FP Type I error	TN

- TP (True Positives): Number of times the model correctly predicts positive samples.
- TN (True Negatives): Number of times the model correctly predicts negative samples.
- FN (False Negatives): Number of times the model incorrectly predicts positive samples as negatives.
- FP (False Positives): Number of times the model incorrectly predicts negative samples as positives.

Fig. 7. Confusion Matrix for a Binary Classifier.

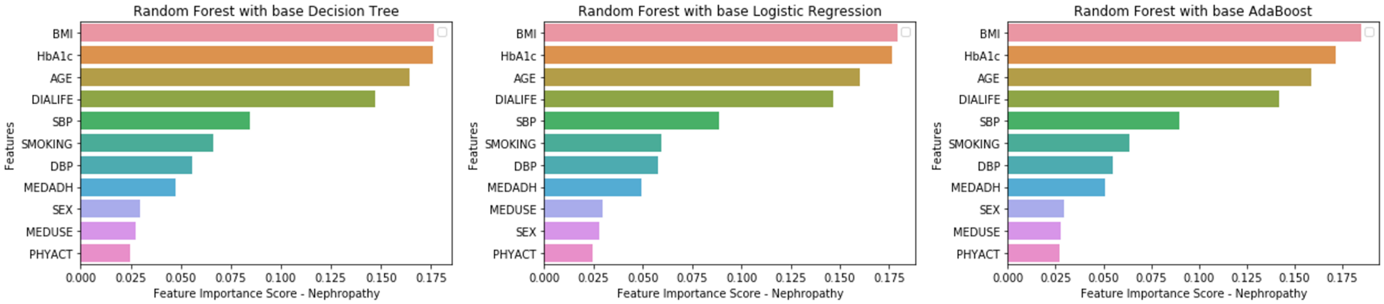


Fig. 8. Feature Importance Scores for Neuropathy.

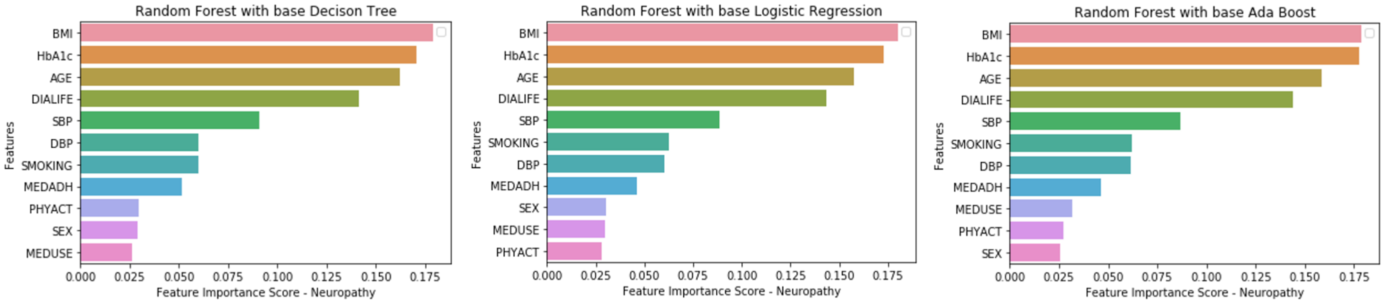


Fig. 9. Feature Importance Scores for Nephropathy.

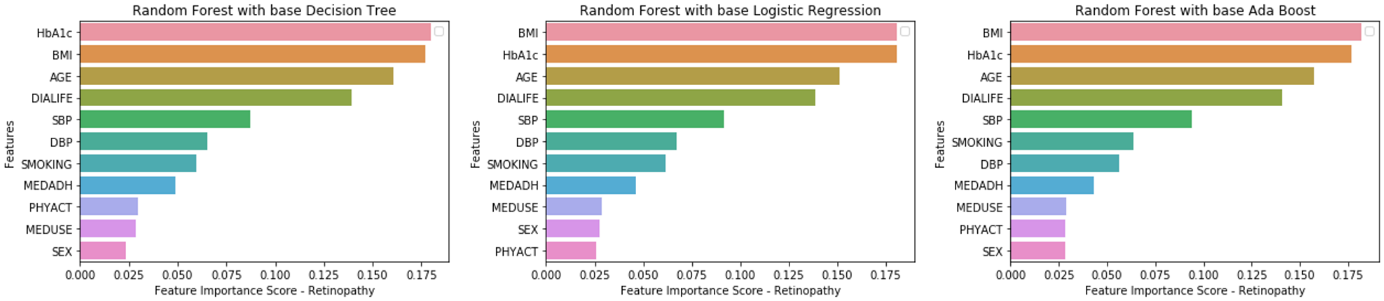


Fig. 10. Feature Importance Scores for Retinopathy.

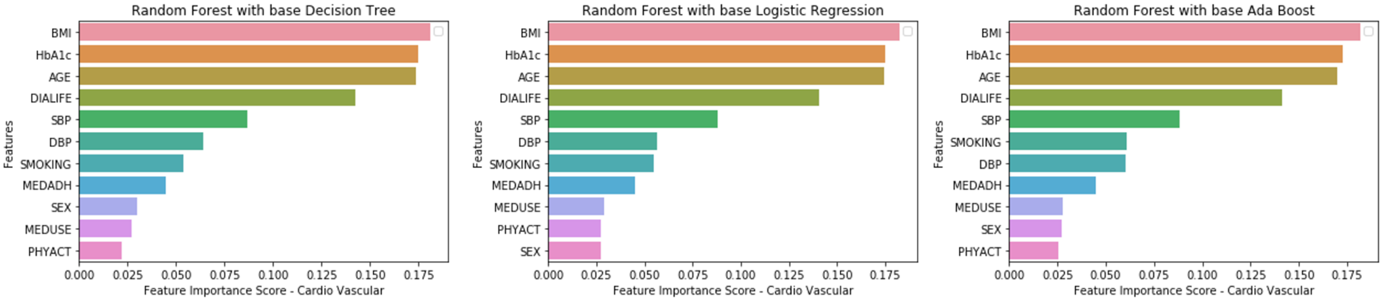


Fig. 11. Feature Importance Scores for Cardio Vascular.

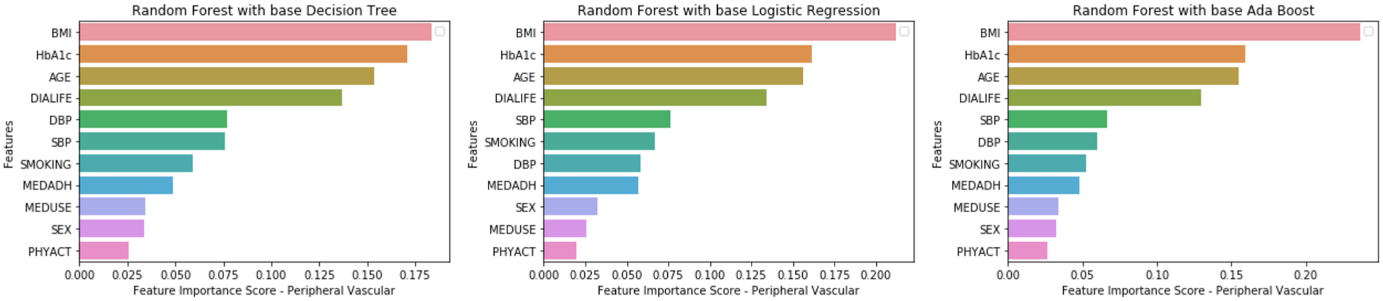


Fig. 12. Feature Importance Scores for Peripheral Vascular.

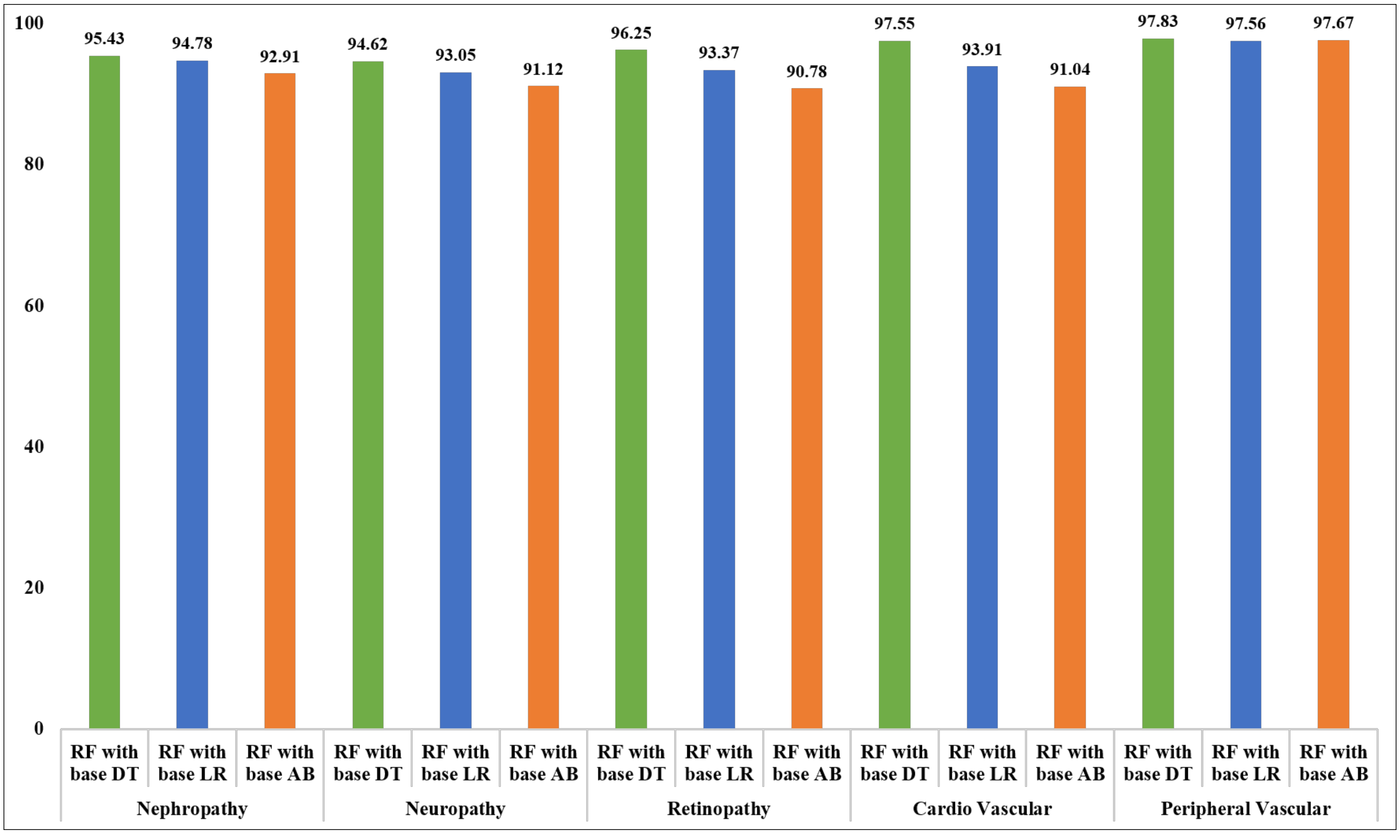


Fig. 13. Accuracy Results for All Complication Modules.

input values (x) are linearly combined using weights values. The output value being modeled is a binary value (0 or 1) rather than a numeric value, which distinguishes it from linear regression. An example LR equation is shown below:

$$y=\frac{1}{1+e^{-\left(w_0+w_1 \times x\right)}} \tag{6}$$

Where y is the predicted output, w_0 is the bias and w_1 is the weight for the single input value (x). Each column in the dataset has an associated w weight that must be learned from the training dataset. LR predicts the probability of the positive class (e.g., a patient has diabetes). For example, if we're predicting whether or not a patient has diabetes or not based on their age, then the positive class could be diabetic, and the LR model could be written as the probability of being diabetic given a patient's age, or more formally:

$$P(x)=P(y=\text{diabetic} \mid x) \tag{7}$$

To get a clear answer, we can snap the probabilities to a binary class value, such as:

$$\begin{aligned} \text{prediction } & =0 \text{ (non-diabetic) IF } p(\text{ diabetic })<0.5 \\ \text{prediction } & =1 \text{ (diabetic) IF } p(\text{ diabetic }) \geq 0.5 \end{aligned} \tag{8}$$

2) Decision Tree: Decision Tree (DT) algorithm is another popular supervised ML algorithm used for solving both classification and regression problems. The goal of using a DT is to learn simple decision rules from training data to create an efficient model that predicts the class of the target variable [34]. DT is a flowchart-like tree structure in which each leaf node corresponds to a class label, and features are represented on the internal node of the tree. The root node is the topmost node in a tree. The DT can be better understood with the help of the algorithm summarized in Fig. 5.

By adopting the above algorithm to the problem of determining whether an instance belongs to class-0 or class-1, we can construct a decision tree by selecting a root node, internal nodes, and leaf nodes, and then defining the class's splitting criteria. For example, we can select the "glucose" feature to be the root node and based on it and other features such as "systolic BP," "Diastolic BP," "Age," and "BMI," we can construct our tree.

The most challenging aspect of the DT algorithm is selecting The root node or first test attribute based on what we will start splitting the data. It is selected based on statistical measures like Information Gain (IG), Gain Ratio, or Gini Index. In this paper, we used the IG measure. IG helps to measure the reduction of uncertainty of a certain feature. It also helps decide which feature is good as a root node. The calculation of $IG(a)$ shows us the formula for the gain in the general case. Let S denote the dataset to be split by creating a new node. Let's suppose that the attribute a can take m values: $a_1, a_2, \ldots, a_m$, and that p_i is the fraction of the objects in S with $a=a_i$. Then, the information gain of a is:

$$IG(a)=H(S)-\sum_{i=1}^m p_i H\left(S \mid a=a_i\right) \tag{9}$$

Here $H(s)$ is the Entropy of the dataset, is a function H of probabilities $p_1, p_2, \ldots, p_n$. Entropy can be thought of as the amount of variance in the data. For binary classification problems, the following formula is used to compute Entropy.

$$H(S)=-\sum_{i=1}^n p_i \log_2 p_i \tag{10}$$

3) Random Forest: Random Forest (RF) is another popular ML technique used to solve regression and classification problems. RF consists of many decision trees. The 'forest' generated by the random forest algorithm is trained using an ensemble method known as bootstrap aggregation (or bagging for short) [35]. Bagging is a technique that combines the predictions from multiple decision trees to make more accurate predictions than any single model. The model's final output is based on the majority of the predictions if the problem is a classification, or the mean of the predictions if the problem is a regression.

4) AdaBoost: Boosting is an ensemble technique for constructing a strong classifier from a collection of weak classifiers. This is accomplished by first creating a model from the training data, followed by the creation of a second model that attempts to correct the errors in the first model. Models are added until the training dataset is perfectly predicted or until the maximum number of models is reached [36]. AdaBoost, shortened for Adaptive Boosting [37], was the first successful boosting algorithm developed for binary classification problems. Decision trees with one level are the most appropriate and widely used algorithm with AdaBoost because these trees are so short and only have one classification decision. They are commonly known as "decision stumps." Weights are assigned to each instance in the training dataset. The initial weight is set as follows:

$$\text{weight}\left(x_i\right)=\frac{1}{M} \tag{11}$$

where x_i is the i'th training instance and M is the number of training instances. To train a single model, a weak classifier is prepared on the training data using the weighted samples. Only binary classification tasks are supported, so each decision stump makes one decision on one input variable and outputs a $+1.0$ or -1.0 value for the first and second class value.

The misclassification rate (E) for the trained model can be calculated as follows:

$$E=\frac{r-M}{M} \tag{12}$$

where r is the number of training instances predicted correctly by the model and M is the total number of the training instances. The opposite of misclassification rate would be accuracy, calculated as:

$$\text{Accuracy}=1-E \tag{13}$$

To take into account the weighting of the training instances, the weighted sum of the misclassification rate is computed as:

$$E=\frac{\sum_{i=1}^n\left(w_i \times p_i\right)}{\sum_{i=1}^n w} \tag{14}$$

where w is the weight for training instance i and p_i is the prediction error for training instance i, which is 1 if misclassified and 0 if correctly classified. For the trained model, a stage value is calculated, which provides a weighting for any predictions made by the model. A trained model's stage value s_v is calculated as follows:

$$s_v = \ln\left(\frac{1-E}{E}\right) \tag{15}$$

The stage weight has the effect of giving more weight or contribution to the final prediction to more accurate models. The training weights are adjusted so that incorrectly predicted instances receive more weight and correctly predicted instances receive less weight. For instance, the weight w of one training instance is updated as follows:

$$w = w \times e^{(s_v \times p)} \tag{16}$$

E. Proposed T2DC Prediction Model

The proposed framework for predicting the risk levels of MIV and MAV complications among T2-DM patients using the RF-based method is shown in Fig. 6. It is broken down into six major steps: a) gathering data from reliable sources; b) cleaning and preprocessing the data; c) dividing the cleaned data into two sets—a training set (80%) and a testing set (20%); d) training the model with the training set; e) evaluating the model performance of the trained model with various base estimators such as DT, LR, and Adaboost models; and f) tuning the model's hyperparameters to see if its accuracy can be improved.

IV. Results and Discussions

This section presents all of the results obtained by the proposed framework, as well as related discussions.

A. Model Performance Evaluation Metrics

Evaluation of the performance of a classification model is based on the number of test samples that the model correctly and incorrectly predicts. The confusion matrix extracts additional information about the performance of a predictive model. It demonstrates which classes are correctly and incorrectly predicted and what types of errors are made. Fig. 7 shows an illustration of a confusion matrix for a binary classifier. The four classification metrics (TP, FP, FN, TN) are calculated, and the confusion matrix compares the model's predicted value to the actual value. The confusion matrix is not exactly a performance metric, but it is used to calculate important classification metrics like accuracy, precision, recall, specificity, and, most importantly, the f1-score, which are used to evaluate the results.

1) **Accuracy:** It is the fraction of correct predictions made by the model.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \tag{17}$$

2) **Precision:** It is the proportion of true positives to all positive predictions made by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \tag{18}$$

3) **Recall:** It is the proportion of actual positives correctly identified by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \tag{19}$$

4) **Specificity:** It is the proportion of actual negatives correctly identified by the model.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \tag{20}$$

5) **F1-score:** It is the proportion of actual positives correctly identified by the model.

$$F1\text{−}score = \frac{2* \text{ precision } * \text{ recall}}{\text{precision } + \text{ recall}} \times 100\% \tag{21}$$

B. Feature Importance Score

Feature importance is a process that involves calculating the score for the input features for a given model — the scores simply represent the "importance" of each feature. A higher score indicates that the specific feature will have a greater impact on the model used to predict a specific class. This can enhance a predictive model's efficiency and effectiveness on the problem. The feature importance scores on T2-DM complications such as nephropathy, neuropathy, retinopathy, cardio vascular, and peripheral vascular modules of the RF model with different base estimators, DT, AdaBoost, and LR, are shown in Fig. 8 to 12, respectively. It is evident from the figures that BMI and HbA1c are the most significant features in the dataset.

C. Model Evaluation Discussion

Table II represents the evaluation of MIV and MAV complications in T2-DM patients. After evaluating the proposed ML models, the decision tree as a base model with RF provides the best performance in all evaluation metrics. The visualization of the accuracy metric for all complication modules is shown in Fig. 13.

D. Risk Prediction Identification

Table III shows the results of T2-DM patients with existing complications, and Table IV shows the predicted complications of the same patients. The existing complications only represent True (1) or False (0) in regard to NEP, NEU, RET, CVD, and PVD. This type of result is not sufficient to assist doctors in identifying further complications. To overcome this, our proposed model can automatically identify the risk levels with respect to low, medium, and high depending upon the severity of T2-DM complications among patients. For instance, T2-DM patient-A has only one NEU complication in existing data. In contrast, the same instance, when evaluated on the proposed risk prediction model, shows a ”low” risk in NEP and a ”medium” risk in NEU. This helps healthcare providers make accurate treatment plans for T2-DM patients so they can give them good clinical care.

V. Conclusion

AI and its applications, such as machine learning (ML) in medical diagnosis and healthcare, have shown enormous promise in recent years, both in terms of improving care and alleviating the enormous strains on the healthcare system. ML-based solutions are revolutionizing diabetes care and helping the medical community gain ground in the fight against the disease. The number of people with diabetes who go undiagnosed can be lowered with the help of ML algorithms that allow for accurate early diagnosis. In this research, we have identified MIV and MAV risk levels of T2-DM complications by proposing a T2DC ML-based prediction model. We have used a decision tree as a base estimator with random forest and obtained better accuracy when compared to other base models. The proposed model achieves 95.43%, 94.62%, 96.25%, 97.55%, and 97.83% accuracies for NEP, NEU, RET, CVD, and PVD complications, respectively. The model can be used as a viable aid in clinical decision-making for practitioners and diabetes educators to improve the quality of life of patients. In the future, we will study how T2-DM and related complications affect pregnant women.

References

[1] S. Ciardullo and G. Perseghin, "Prevalence of elevated liver stiffness in patients with type 1 and type 2 diabetes: A systematic review and meta-analysis," *Diabetes Research and Clinical Practice*, p. 109981, 2022.

[2] X. Ding, S. Rong, Y. Wang, D. Li, L. Wen, B. Zou, D. Zang, K. Feng, Y. Liang, F. Wang *et al.*, "The association of the prevalence of depression in type 2 diabetes mellitus with visual-related quality of life and social support," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 15, p. 535, 2022.

[3] N. Sambyal, P. Saini, and R. Syal, "A review of statistical and machine learning techniques for microvascular complications in type 2 diabetes," *Current Diabetes Reviews*, vol. 17, no. 2, pp. 143–155, 2021.

[4] Q. Xu, L. Wang, and S. S. Sansgiry, "A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning," *J. Med. Artif. Intell*, vol. 3, no. 6, 2020.

[5] E. Adua, E. A. Kolog, E. Afrifa-Yamoah, B. Amankwah, C. Obirikorang, E. O. Anto, E. Acheampong, W. Wang, and A. Y. Tetteh, "Predictive model and feature importance for early detection of type ii diabetes mellitus," *Translational Medicine Communications*, vol. 6, no. 1, pp. 1–15, 2021.

[6] R. Wang, Z. Miao, T. Liu, M. Liu, K. Grdinovac, X. Song, Y. Liang, D. Delen, and W. Paiva, "Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records," *Journal of Clinical Medicine*, vol. 10, no. 7, p. 1473, 2021.

[7] M. Maniruzzaman, M. M. Islam, M. J. Rahman, M. A. M. Hasan, and J. Shin, "Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 5, p. 102263, 2021.

[8] L. Muhammad, E. A. Algehyne, and S. S. Usman, "Predictive supervised machine learning models for diabetes mellitus," *SN Computer Science*, vol. 1, no. 5, pp. 1–10, 2020.

[9] A. Al Bataineh and A. Jarrah, "High performance implementation of neural networks learning using swarm optimization algorithms for eeg classification based on brain wave data," *International Journal of Applied Metaheuristic Computing (IJAMC)*, vol. 13, no. 1, pp. 1–17, 2022.

[10] A. A. Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *International Journal of Machine Learning and Computing*, vol. 9, no. 3, pp. 248–254, 2019.

[11] A. Al Bataineh, D. Kaur, and S. M. J. Jalali, "Multi-layer perceptron training optimization using nature inspired computing," *IEEE Access*, vol. 10, pp. 36 963–36 977, 2022.

[12] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, and A. Alhumam, "Investigating health-related features and their impact on the prediction of diabetes using machine learning," *Applied Sciences*, vol. 11, no. 3, p. 1173, 2021.

[13] A. Allen, Z. Iqbal, A. Green-Saxena, M. Hurtado, J. Hoffman, Q. Mao, and R. Das, "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus," *BMJ Open Diabetes Research and Care*, vol. 10, no. 1, p. e002560, 2022.

[14] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Applied Intelligence*, vol. 52, no. 3, pp. 2411–2422, 2022.

[15] M. Rashid, M. Alkhodari, A. Mukit, K. I. U. Ahmed, R. Mostafa, S. Parveen, and A. H. Khandoker, "Machine learning for screening microvascular complications in type 2 diabetic patients using demographic, clinical, and laboratory profiles," *Journal of Clinical Medicine*, vol. 11, no. 4, p. 903, 2022.

[16] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *International journal of environmental research and public health*, vol. 18, no. 6, p. 3317, 2021.

[17] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 103 737–103 757, 2021.

[18] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A machine learning approach to predicting diabetes complications," in *Healthcare*, vol. 9, no. 12. MDPI, 2021, p. 1712.

[19] G. Naveen Kishore, V. Rajesh, A. Vamsi Akki Reddy, K. Sumedh, and T. Rajesh Sai Reddy, "Prediction of diabetes using machine learning classification algorithms," *Int J Sci Technol Res*, vol. 9, no. 01, pp. 1805–1808, 2020.

[20] L. C. Jung, H. Wang, X. Li, and C. Wu, "A machine learning method for selection of genetic variants to increase prediction accuracy of type 2 diabetes mellitus using sequencing data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, no. 3, pp. 261–281, 2020.

[21] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[22] M. S. Islam, M. K. Qaraqe, and S. B. Belhaouari, "Early prediction of hemoglobin alc: A novel framework for better diabetes management," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 542–547.

[23] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[24] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine learning methods to predict diabetes complications," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 295–302, 2018.

[25] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. IEEE, 2018, pp. 291–295.

[26] Y. Fan, E. Long, L. Cai, Q. Cao, X. Wu, and R. Tong, "Machine learning approaches to predict risks of diabetic complications and poor glycemic control in nonadherent type 2 diabetes," *Frontiers in Pharmacology*, vol. 12, p. 1485, 2021.

[27] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia computer science*, vol. 161, pp. 449–457, 2019.

[28] B. Sudharsan, M. Peeples, and M. Shomali, "Hypoglycemia prediction using machine learning models for patients with type 2 diabetes," *Journal of diabetes science and technology*, vol. 9, no. 1, pp. 86–90, 2014.

[29] B. Vamsi and D. Bhattacharyya. (2021) Micro and macro

vascular complications in type_ii diabetes. [Online]. Available: https://data.mendeley.com/datasets/dsjcb6pyd8/1

[30] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[31] A. Al Bataineh and D. Kaur, "Immunocomputing-based approach for optimizing the topologies of lstm networks," *IEEE Access*, vol. 9, pp. 78 993–79 004, 2021.

[32] A. Al Bataineh and S. Manacek, "Mlp-pso hybrid algorithm for heart disease prediction," *Journal of Personalized Medicine*, vol. 12, no. 8, p. 1208, 2022.

[33] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[34] J. R. Quinlan, "Probabilistic decision trees," in *Machine Learning*. Elsevier, 1990, pp. 140–152.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[36] J. Brownlee, *Machine learning algorithms from scratch with python*. Machine Learning Mastery, 2016.

[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.