Grenze International Journal of Engineering and Technology, June Issue

Class Disproportion using a Variety of Machine Learning Approaches to Predict Clinical Risk Factors for Diabetes

Mr. Ranjith Bhat[1], Dr. Aruna[2], Ms. Shruthi S[3], Ms. Veena M S[4] and Dr. Raghu N[5]
[1]Assistant Professor, Department of Robotics and AI NMAM Institute of Technology, Nitte, India.
Email: n.raghu@jainuniversity.ac.in
[2] Department of EEE, Nitte Meenakshi Institute of Technology, Bangalore, India.
Email: arunakiranka@gmail.com
[3-4] Department of CSE, R R institute of Technology, Bangalore, India.
Email: shruthiraju9184@gmail.com, veenshekar009@gmail.com
[5] Department of EEE, JAIN (Deemed-To-Be-University), Bangalore, India.
Email: ranjithbhat@gmail.com

***Abstract*— Diabetes is the most prevalent and quickly spreading disease that shortens life expectancy. It affects a great number of people annually from all age groups. posing a strong impact rate, it makes the first diagnosis more significant. Diabetic carries additional difficult consequences such as cardiovascular disease, renal failure, stroke, harm to important organs, etc. Earlier Diabetes diagnosis lowers the risk that it may progress into a severe and ongoing condition. Determining the frequency of diabetes in medical analysis is facilitated by the identification and assessment of risk factors for certain spinal features. Future consequences are less likely when diabetes is detected early on and its prevalence is measured. This study used the combined NHANES dataset from 1999–2000 to 2015–2016. Its goals were to use multiple supervised machine learning algorithms to identify abnormalities and to analyse and determine possible risk variables connected to diabetes by means of ANOVA and logistic regression. The experimental results, which corrected for class imbalance and outlier issues, indicate that blood-related diabetes, age, cholesterol, and BMI are the main risk factors for developing diabetes. Along with this, the random forest classification approach yielded the greatest accuracy score of 90.**

***Index Terms*— Safety Factors Analysis, Categorization, Exception, Class Difference, Logistic Regression, Machine Learning.**

I. INTRODUCTION

One of the most common and quickly spreading diseases in the world today is diabetes. Diabetes mellitus, another name for diabetes, is a metabolic condition. It considerably raises blood sugar levels [1]. Diabetes raises the risk of further long- term issues, such as cardiovascular disease, kidney failure, and heart attacks. Patients with diabetes who are older also experience various forms of serious damage to their blood vessels, essential organs, and nerves [2]. Individuals between the ages of 20 and 80 are most vulnerable to the effects of diabetes. By 2045, there will

Grenze ID: 01.GIJET.10.2.177_1

be 700 million more adults afflicted by this illness than there are currently (465 million) [3]. Diabetes has a very high death rate since it is often associated with other complex medical conditions. Diabetes was a contributing factor in over 4.2 million deaths [1][2]. Over the past 20 years, the number of diabetics affected has doubled worldwide [3].

The impact and death rates from diabetes illnesses are rising quickly every day. Due to the delay in diagnosis, the effects of diabetes become increasingly terrifying. Early diagnosis of diabetes helps reduce the serious long-term effects of this condition. Diabetes was among the most researched topics among researchers due to its link with other chronic, long-standing diseases. The examination of diabetes became more difficult and sophisticated due to nonlinear, complex, outliers, and other medical data kinds [1]. In recent times, health informatics systems have become indispensable in the identification and tracking of several illnesses and ailments. Various machine learning-based algorithms can greatly improve the diagnosis of impacted instances and assess the patients' early risk factors to determine the disease's severity. Several machine learning techniques have been developed by researchers in recent studies to categories people with diabetes. Numerous statistical analyses were conducted in order to identify prospective diabetes risk variables [4]. Several machine learning techniques were applied in those studies. However, some studies' data preparation is deficient [1]. Inadequate handling of class imbalance, outliers, etc., can have a big impact on the final result. In order to discover potential clinical risk variables and their impact on diabetes [5], as well as to recommend a system that can accurately detect diabetes patients, this study integrated statistical and machine learning techniques. To determine which risk factors were most significantly connected with diabetes, an analysis of the risk factors of several independent variables was conducted. Diabetes patients are categorized using a variety of machine learning techniques in addition to risk analysis. Our suggested system's overall workflow is displayed in figure 2 and figure 3 depicts the suggested architecture for the ANN in Methodology section. The extreme numbers were swapped out for the median values to deal with the outliers. Class weights were assigned in order to address the issue of [6] class imbalance. Analysis of variance (ANOVA) and statistical logistic regression were used to analyse the risk factors. The characteristics were given in many classifications of algorithms after the exception and class difference problem were resolved. The classification presentations were analyzed and contrasted with further state-of-the-art methods, as indicated in Tables I and II further below.

II. Literature Survey

The study on Predicting Type 2 Diabetes Mellitus: A Comparison between Logistic Regression, Decision Trees, and Multi- layer Perceptrons" examined the accuracy of three different methods of type 2 diabetes mellitus prediction: multilayer perceptrons, decision trees, and logistic regression. The results showed that decision trees performed better than other approaches, yielding higher accuracy and improved predictive power, because of their capacity to manage non-linear correlations and interactions among risk factors.

The Pima Indian Diabetes dataset was used in Machine Learning Approach for Diabetes Risk Prediction Using Pima Indian Diabetes Dataset," to predict [7] diabetes risk using a variety of classification techniques. Support Vector Machine (SVM) was shown to be the most accurate algorithm tested, suggesting that it is a suitable approach for detecting persons who are at risk.

The focus of Predicting the Onset of Diabetes Mellitus Using Machine Learning was on making such predictions. According to this study, k-Nearest Neighbours (k-NN) and logistic regression are useful methods for determining risk factors linked to the onset of diabetes. These results highlight the importance of k-NN and logistic regression in identifying risk factors early on, which helps with early diagnosis and intervention.

The study on Predicting Diabetes Mellitus Using SMOTE and Ensemble Techniques" investigated the application of ensemble methods and the Synthetic Minority Over-sampling Technique (SMOTE) for the prediction of diabetes mellitus. Of these, the ensemble method performed better [8] in terms of prediction, especially when using Random Forest. This demonstrates how ensemble approaches work well in conjunction with SMOTE to alleviate class imbalance and produce more reliable diabetes prediction models. An in-depth analysis of machine learning techniques for diabetes prediction was provided in Diabetes Mellitus Prediction Models using Machine Learning: A Review on Current Approaches. The review found that k-Nearest Neighbours, Random Forest, and Support Vector Machines were effective techniques. Random Forest produced the greatest outcomes in certain instances. The main conclusion is that machine learning models, such k-NN, SVM, and Random Forest, have potential for diabetes prediction. In addition, correcting the imbalance in classes is an important part of improving model performance.

All of these observations add to our understanding of efficient machine learning techniques for estimating the risk of diabetes. Important contributions to enhancing the precision and dependability of diabetes prediction models are made by decision trees, SVM, logistic regression, k-NN, ensemble approaches, and class imbalance handling strategies.

III. DATASET

A. About the Dataset: The National Health and Nutrition Analysis Survey provided the dataset for this study (NHANES). For this investigation, the assembled version NHANES data from 1999–2000to 2015–2016 was used. NHANES is a health and nutrition calculation programme that handles a variety of data typesand is based on the US population. The dataset employed in this study is an aggregate of laboratory, survey, demographic, and examination data from 1999–2000 to 2015–2016. Several studies have also made use of the dataset [4]. Responses from 37079 individuals with 51 distinct categories of attributes are included in the sample. [10] For the purpose of identifying diabetes disorders based on their impacts, a total of twenty five independent variables were chosen. Diabetes is regarded as a dependent variable for this study. This dataset includes 32227 normal patients and 4852 diabetic cases, for a total of 37079 responders. The dataset includes an imbalance in classes basedon the distribution of diabetes patients' classes; this issueis resolved by allocating class weights during training using various machine learning strategies.

B. Exploratory Data Analysis (EDA): The dataset encloses the evidence from 37079 respondents and roughly 51 unique factors. Gender, the proportion of family income to poverty, weight, systolic and diastolic white blood cells, albumin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), cholesterol, creatinine, glucose, gamma-glutamyl transferase (GGT), iron, lactate dehydrogenase (LDH), phosphorus, bilirubin, total cholesterol, high-density lipoprotein (HDL), glycohemoglobin, vigorous work, moderate work, health insurance, coronary heart infection. Diabetes is considered a dependent variable in this study. For this study, twenty five distinct variables were selected based on their significance and effect on individuals with diabetes [11]. Certain variables (such as the ratio of family income to poverty, the pulse rate in the 60s, the yearly family income, lymphocytes, monocytes, eosinophils, basophils, mean cell volume, etc.) are not taken into consideration for additional analysis. The computer language Python was utilized in this study to conduct a variety of analysesand experiments [12]. Figure 1 shows the removal of all the un contributed outliers on the basis of the Z-score and a box-plot plotting of all the different conditions of diabetes mentioned here as B2A*, where * being values 1,2,3,4.

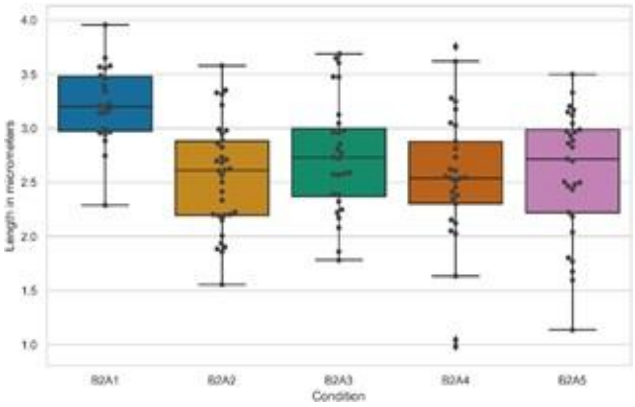

Figure 1. Boxplot after handling outliers based on Z score.

IV. DATA PREPROCESSING

A. Managing the Outlier Using the Z Score: Data points that deviate from other data points in the dataset distribution are referred to as outliers [13]. Outliers are anomalous or uncommon values found in a specific data distribution.Outliers have a major impact on the performance of machine learning algorithms and statistical decision-making. Figure 1 displays the boxplot visualizations of the dataset's chosenindependent variables that were created for outlier analysis.It is evident from figure 1 that outliers have an impact on the dataset. If they are not handled, it could affect how well algorithms work and how risk variables are analysed. TheInterquartile Range (IQR) approach was used in this studyto identify outliers [14].

The interquartile values span the first through the third quartile. The IQR is represented mathematically as follows:

$$IQR = Q3 - Q1 \quad (1)$$

Where the first and third quartiles are denoted, respectively, by Q3 and Q1. In order to identify the outliers and indicate the greatest and lowest value ranges, the lower and upper fences were computed. Outliers are values that deviate from both the lowest and maximum allowable range [15], as illustrated in Figure 1. The lower and upper fence's mathematical representation is as follows:

$$Upperfence = Q3 + (1.5 * IQR) \quad (2)$$

$$Lowerfence = Q1 - (1.5 * IQR) \quad (3)$$

Systolic, diastolic, weight, body mass index, hemoglobin, albumin, ALP, AST, ALT, cholesterol, GGT, iron, LDH, HDL,and glycohemoglobin are the 15 variables in total that are impacted by outliers. To deal with the outliers, the extreme values were first replaced with their median values and then eliminated entirely. Therefore, the median values of the corresponding variables displayed in figure 1 are used to replace the outliers.

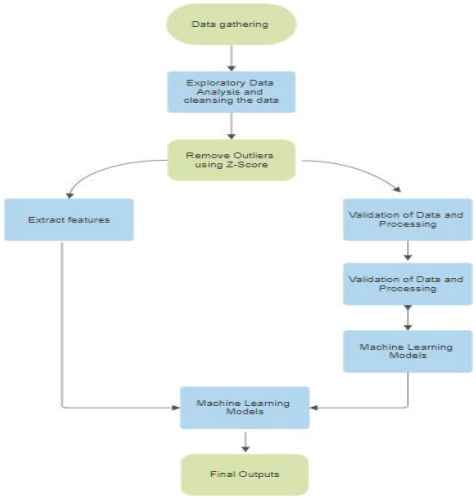

Figure 2. Flow of the proposed system.

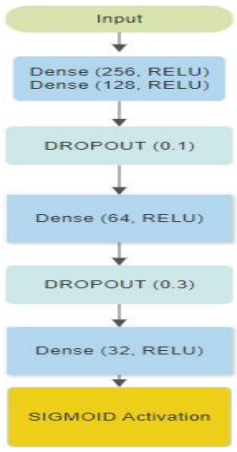

Figure 3. Proposed Artificial Neural Network architecture.

B. Initialization of Weights: The dataset has 37079 respondents, of which 4852 have diabetes and 32227 are normal individuals. This suggests that there is a problem with class imbalance. Using the scikit-learntool, class weights (Class 0: 0.52, Class 1: 0.47) are assignedto the classes in order to address the issue of class imbalance [16].

C. Data Standardization: The process of standardization involves rescaling the value distribution to maintain a mean of 0 and a standard deviation (std) of 1. Additionally, outliers have less of an impact [9].In this study, the scikit-learn tool was used to accomplishdata standardization following the management of the dataset's outliers [17].

IV. METHODOLOGY

A. Feature Importance: There are 51 independent variables in the dataset that was used in this study. However, a univariate feature collectionapproach and statistical logistic regression method were used for this research to account for 25 independent variables, representing their impacts on patients with diabetes [18]. ANOVA F-scores for each variable verified the results ofrisk factors found by statistical logistic regression. Also, a comparison and validation of the resulting risk variables are made with the results of other studies [1].

Analysis of Variance (ANOVA): There are 25 independent variables in the dataset used for this study. An algorithm for feature selection was used to identify the best features.Methods for feature selection also improve the accuracy of predictions. The analysis of variance (ANOVA), which is also employed in other studies [10], was used in this investigation as a feature selection method. ANOVA is a straightforward and effective statistical technique for analyzing the means of one or more groups. It also establishes the degree to which the groups differ from one another.

1)Logistic Regression (LR): Previous studies on diabetes patients' risk factor identification have employed logistic

re- gression [20]. Due to its binary structure, LR is one of the many popular supervised machine learning algorithms. The possible risk factors connected to diabetes patients are also examined in this study using statistical LR. A logit is thedependent variable in LR, and the equation is as follows

$$\log(x) = \log \frac{x}{x-1} = B_0 + B_1Y_1 + \ldots + B_YY_Y \quad (4)$$

2) Risk Factors Analysis using Logistic Regression: The logistic regression representation is employed in this study to explore statistical data. The relative risk variables in our dataset were measured and identified using the stats models programme [11]. Table I displays the p-value, confidence interval (C.I.), and odds ratio (O.R.) for each variable that was generatedto examine the relative risks of the independent variables with respect to the outcome variable. These results indicatethe risk and impacts of individual qualities on the dependent variable. The p-value is a useful tool for representing statisticalsignificance. A p-value might be anywhere between 0 and 1 [19]. For statistical significance in this study, a p-value of less than 0.05 is considered acceptable. The most important risk factors, in addition to those linked to diabetes disease, include age, diastolic, cholesterol, and blood-related diabetics (Table I). The relative risk connected to the outcome variable is also examined using the odds ratio (OR).

A. Cross Validation: The dataset is divided into two groups using a data partitioning technique called cross-validation (CV): train data and test/validation data. The cross-validation method in this study utilized the Shuffle Split operation, which generates an arbitrary number of distinct train and validation splits. The samples in this method are divided into train and test/validation sets by shuffling them. Having control over the quantity of data samples on both sides of the train/validation set and the number of iterations is the reason this method was chosen. The dataset was divided into 80% training set and 20% set/validation set using a cross-validation procedure. The data was analyzed using CV/split values of 3, 5, and 10 for the train and validation, respectively, using the scikit-learn program [8].

B. Prediction Models: Several machine learning predictive models were employedin this study to assess the effectiveness of the suggested approach. The proposed architecture and the flow of work have been represented below in the Figure 3. where x is the likelihood thata patient has diabetes when y = 1, and 1-x is the probability that a patient does not have diabetes when y = 0. The statisticallogistic regression method, which determines potential risk factors connected with diabetes, was also used to identify the odd ratio and confidence intervals. The features were chosen and arranged in order of their corresponding p-values. P-values less than 0.05 were taken into account in this study in order to determine the specific risk factors. Table I displays the risk factors along with their p-value, odd ratio, and confidence range.

1) Support Vector Machine (SVM): Previous research on diabetes risk factor detection have also employed support vector machines (SVM) [13]. SVM is a well-liked supervised learning technique that may be used to both solve classification issues and look for hidden patterns. SVM's ability to locate a hyperplane also makes it recognized as a decision boundary function. This method converts the original train data into a higher dimensional space under the supervision of nonlinear mapping. Afterwards, SVM distinguishes between patterns be- longing to various classes [14]. The hyperplane's mathematicalequation is:

$$W.Y + p = 0 \quad (5)$$

In this case, b denotes the scalar data and W the weighted vector. One of the classification techniques employed in this study is the Radial Basis Function (RBF) kernel of SVM.

2) Random Forest (RF): Random forests are an ensemble learning method for problems like regression and classification. They are often referred to as random decision forests. During the training phase, they construct a huge number of decision trees, and the resultant class indicates the mean/average prediction (regression) or the mode of the classes (classification) of the individual trees. Previous research on diabetes diagnosis has also employed Random Forest classification [1]. Regression and classification issues are two scenarios in which the random forest can be used. The decision tree'sproperties serve as the foundation for this enhanced versionof the bagging technique, which is quick and resistant to overfitting [15].

3) Decision Tree (DT): The investigations also employclassification based on decision trees [1]. DT is a well-liked supervised learning technique that may be applied to bothregression and classification models. It creates tree-structure- based regression or classification models. Using this method, the input features are

4) Artificial Neural Network (ANN): Several studies that predict diabetes also use artificial neural networks (ANN) [2]. In this study, dropout layers were included in addition toseveral dense layers to avoid overfitting. In order to address the issue of class imbalance, the dataset was trained usingclass weights. Figure 3 depicts the suggested

architecture of the ANN.

Ensemble: The ensemble method of assessing the model's performance is incredibly effective [16]. In this study, the hard-voting approach was used for ensemble [17]. Allof the earlier techniques were combined with the hard-voting method in this study.

V. Results And Discussion

A. Evaluation Metrics

Several statistical measures were utilized in this study as assessment metrics to measure the risk features and the impacts of the different variables on an independent variable, including the odd ratio (OR), P-value, and confidence interval (CI). In addition to this, the efficiency of our suggested model was assessed using assessment metrics such as accuracy and AUC score.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \qquad (6)$$

Where the values for True Positives, True Negatives, False Positives, and False Negatives are represented, respectively, by the symbols TP, TN, FP, and FN [18]. The AUC score, which illustrates the sensitivity versus 1 - specificity and suggests the total efficiency of the model, is the product of the True Positive (TP) and False Positive (FP) rates.

TABLE I. LOGISTIC REGRESSION USING INDIVIDUAL RISK FACTORS AND THEIR P-VALUE, ODD RATIO, AND CONFIDENCE INTERVALS

Features	P value	Old Ratio(OR)	5% CI	95% CI
Age	0.00	1.055761	1.047528	1.0641
Diastolic	0.00	1.023792	1.013933	1.03776
Cholesterol	0.00	1.032076	1.020663	1.04361
Blood-Rel-Diab	0.00	1.063944	1.052669	1.0753
Hemoglobin	0.0004	1.028249	1.011516	1.0411
BMI	0.0050	1.173713	1.55784	1.2919
AST	0.0013	1.021564	1.008342	1.0349
HDL	0.0022	1.018803	1.006727	1.0310
Gender	0.0048	1.023693	1.007169	1.0404
Phosphorus	0.0092	0.985706	0.975092	0.9964
Systolic	0.0299	0.986081	0.973684	0.9986
Protein	0.0306	0.986909	0.975190	0.9987
ALT	0.0309	0.985280	0.972104	0.9986

B. Statistical Analysis

The variables in this study that have a p-value of less than 0.05 were chosen for examination based on their statistical significance. The most important risk factors linked to diabetesclass include age, diastolic blood pressure, cholesterol, blood- related diabetes, BMI, etc. Figure 3 also displayed the feature importance following the application of Analysis of Variance (ANOVA) to handle the outliers. The chosen risk variablesare validated by figure 3 based on the F-score. The most important risk variables for diabetes include age, blood-relateddiabetes, cholesterol, and BMI, according to the intersectionof statistical logistic regression and ANOVA methodologies. By contrasting these risk factor findings with those from otherinvestigations, they were verified and validated [21]. A larger odd ratio is linked to an increased likelihood of falling intothe diabetes category. In a similar vein, a smaller odd ratio is associated with a lower likelihood of falling into the diabetic category. Table I identifies blood-related diabetes variables, age, diastolic blood pressure, cholesterol, BMI, and gender as high-risk factors for the development of diabetes.

C. Results

The comparison of accuracy and AUC scores with various CV values is shown in Table II. Three, five, and ten split/cv values were used to test the dataset. In this study, a number of classification techniques were employed, including ensemble approach, support vector machine, logistic regression, random forest, decision tree, and artificial neural network. The assessment metrics listed in Table II have undergone considerable modifications. The randomized forest classification approach (RF) outperformed the other classification techniques in terms of efficiency (90%) and AUC score (0.98).

TABLE II. PERFORMANCE AND AUC SCORE COMPARISON USING DIFFERENT CV VALUES

Methods	Evaluation Metrics	CV 3	CV 5	CV 10
ANN	Acc.	.80	.80	.82
	AUC	.80	.80	.82
DT	Acc.	.85	.86	.86
	AUC	.87	.95	.96
LR	Acc.	.87	.88	.88
	AUC	.88	.83	.84
RF	Acc.	.90	.90	.90
	AUC	.89	.97	.98
SVM	Acc.	.88	0.88	.89
	AUC	.89	.90	.90
Ensemble	Acc.	.87	.88	.90
	AUC	.86	.89	.90

V. CONCLUSION

The primary goals of this study are to determine and examine probable clinical risk aspects for diabetes and to providean effective system for categorizing diabetic patients based ona variety of characteristics. In this study, 37079 observationsof patients with various characteristics were examined in total.Rather than eliminating the outliers entirely, the extreme numbers were replaced by median values to address the outliers. The issue of class imbalance was resolved during training by allocating class weights. ANOVA and logistic regression wereused to find and examine the possible clinical risk factors. Therefore, the most important risk factors for diabetes are revealed to be age, blood-related diabetes, cholesterol, and BMI characteristics. The processed dataset is subjected to many classification algorithms, yielding high accuracy andAUC values that are compared with other cutting-edge techniques. In our suggested approach, the random forest algorithmproduced the best accuracy (.90) and AUC score (.98). Evenif these scores are better than those of other efforts that have already been done, they could still be raised by using additional advanced techniques.

REFERENCES

[1] Alzboon, Mowafaq Salem, et al. "Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods." International Journal of Online & Biomedical Engineering 19.15 (2023)..

[2] S. M. Moonsamy, D. Parikh, and A. Ramphul, ”Pre- dicting diabetes using machine learning techniques: A review,” Journal of Medical Systems, vol. 44, no. 5, p. 93, 2020.

[3] D. C. Kwon, M. W. Kim, and B. C. Lee, ”Diabetes prediction model using machine learning and big data,” Healthcare Informatics Research, vol. 27, no. 2, pp. 154- 159, 2021.

[4] S. S. Meena, R. P. Singh, and P. K. Sharma, ”Predictive modeling for diabetes using different machine learning techniques,” Procedia Computer Science, vol. 85, pp. 127-134, 2016.

[5] M. S. Dhillon, A. S. Sahani, and A. Mittal, ”Predic- tion of diabetes using machine learning algorithms,” in 2018 International Conference on Sustainable Energy, Electronics, and Computing Systems (SEEMS), 2018, pp. 236-241.

[6] Y. Chen, H. Q. Huang, J. Xu, and Z. R. Xiao, ”Diabetes prediction with machine learning techniques,” in 2018 37th Chinese Control Conference (CCC), 2018, pp. 5658-5663.

[7] I. S. Andrikopoulou, I. Koutsouris, and D. K. Iakovidis, ”Machine learning for the prediction of diabetes,” in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 3367-3370.

[8] Y. Lin, L. Zhang, and Z. Lai, ”Predicting diabetes with machine learning models: A systematic review,” Journal of Healthcare Engineering, vol. 2021.

[9] A. Al-Mardini, I. M. Bani-Hani, and M. M. Al- Khateeb, ”A predictive model for diabetes using data mining techniques,” in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1-5.

[10] A. Almarashi, S. Al-Mejibli, and H. J. Almarashi, ”A review of machine learning methods for predicting the onset of diabetes mellitus,” Journal of Healthcare Engineering, vol. 2020.

[11] S. Mahmood, R. K. B. Rahman, and M. A. Ahmed, ”Diabetes prediction using machine learning and ad- vanced analytics,” in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018, pp. 1-6.

[12] H. S. Magar, S. Y. Bahutule, and K. P. Yadav, ”Diabetes prediction using machine learning techniques,” in 2019 4th International Conference on Computing, Communi- cation and Security (ICCCS), 2019, pp. 1-5.

[13] S. M. Hoque, F. M. Al-Shorbaji, and F. S. B. Jammal, ”Predictive modeling of diabetes using machine learning algorithms,” in 2018 7th International Conference on Software and Computer Applications (ICSCA), 2018, pp. 143-147.

[14] G. E. Djibril, S. S. Kabore, and I. P. Ouattara, ”Diabetes prediction with machine learning techniques,” in International Conference on Wireless Networks and Mo- bile Communications (WINCOM), 2018, pp. 97-102.

[15] Fan Y, Long E, Cai L, Cao Q, Wu X, Tong R. Machine Learning Approaches to Predict Risks of Diabetic Complications and Poor Glycemic Control in Nonadherent Type 2 Diabetes. Front Pharmacol. 2021.
[16] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, ”An efficient convolutional neural network for coronary heart disease prediction,” Expert Systems with Applications, p. 113408, 2020.