Date of publication DEC-31, 2025, date of current version DEC-14, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.3.4342

Classification of Diabetes using Multinomial Naive Bayes, Logistic Regression, and Multi-Layer Perceptron Algorithms

EMAD MAJEED HAMEED[1,2], HARDIK JOSHI[1], HAYDER J. ALHAMDANE[3], MUDHAR A. AL-OBAIDI[4]

[1]Gujarat University, Ahmadabad, Gujarat, India
[2]Technical Institute of Baquba, Middle Technical University, Diyala, Iraq
[3]College of Applied Arts, Middle Technical University, Bagdad 10074, Iraq
[4]Technical Instructor Training Institute, Middle Technical University, Baghdad 10074, Iraq

Corresponding author: Mudhar A. Al-Obaidi (e-mail: dr.mudhar.alaubedy@mtu.edu.iq).

ABSTRACT Diabetes is an ongoing condition in which a human being's blood glucose levels increase to unacceptably high levels. For the purpose of organizing the required treatments and avoiding the development of more severe diseases that this disease may bring on, diabetes should be detected as early as possible. In this study, diabetes is classified through the use of different models and to determine the most appropriate model that can be used for this problem. In this study, Logistic Regression, Multinomial Naive Bayes, And Multi-Layer Perceptron Algorithms are utilised as classification models. The Indian Pima Data Set is utilised to test these techniques. The preprocessing steps used in this study involve working with the noisy data, scaling of data using normalization, processing imbalanced data using the SMOTE approach, and using sequential backward selection technique (SBS) for features selection. The classification performances of techniques Logistic Regression, Multinomial Naive Bayes, and Multi-Layer Perceptron obtained by dividing dataset into 80% training dataset and 20% testing dataset are 74.5%, 78%, and 62%, respectively. This study has specifically solved the issues of under fitting and overfitting.

KEYWORDS Diabetes; Prediction; Multinomial Naive Bayes; Logistic Regression; MLP.

I. INTRODUCTION

Diabetes is a condition that arises from a deficiency, ineffectiveness, or insufficient production of the hormone insulin in the body. Chronic complications cause disruptions in carbohydrate metabolism and raise blood glucose levels. If diabetes is not treated, it can lead to numerous complications for the patient, including intense thirst, intense hunger, and frequent urination. Inadequate precautions and uncontrolled blood sugar have a negative impact, particularly on the vessels. The main organs and tissues that sugar damages permanently include our heart, brain, and leg vessels, as well as our eyes, kidneys, and nerve endings [1, 2]. Therefore, early diagnosis of diabetes is vital to prevent many damages. Medical studies have shown that the pathology of diabetes has recently gotten worse and does not typically stop. There are currently 537 million diabetics worldwide, and just in 2021, 6.7 million people died from the disease [3]. The detection of diabetes can be made by human health experts as a result of manual examinations or by examining blood samples taken from patients with the help of a medical device in a laboratory environment. However, since diabetes is a disease that progresses without showing many symptoms, it may not be clearly diagnosed even by doctors who are experts in their field [4].

With technological developments, it has become possible to diagnose many diseases using intelligent and learning approaches. In this way, diagnosis of diseases and reporting of relevant examinations are completed in a shorter time, and as a result, the time spent by patients in the healthcare institution is reduced [5]. Nowadays, large investments are being made in smart hospital projects in many countries. This application both relieves the density in healthcare institutions and reduces the amount of labor required by automating the system. Approaches based on machine learning and data mining are of great interest for the detection, management, and other related clinical treatment of diabetes. The early diagnosis diabetes disease is greatly helped by computer-aided expert systems built on machine learning [6, 7]. This paper aims to apply data mining techniques to early detect diabetes.

When compared to the studies in the literature, thanks to the variety of methods used in the exploratory data analysis stage of this study, resolving the problem of missing values,

determining and fixing of the outliers values, and imbalances in the dataset were determined, making the dataset more suitable for classification models. When there is a large gap in the data, normalization is used to rescale the data so that it falls within a smaller range. It serves to enhance the efficiency and reliability of machine learning model [8].

In the second part of this study, a literature summary is given by referring to the studies in the literature. In the 3rd section, the approaches used within the scope of the study, the data set and the preprocessing carried out until the stage of making the data set suitable for the models to be applied are discussed. In the 4th section, the experimental results of the study are mentioned. Finally, in the 5th section, the study is concluded by giving the results of the study.

II. LITERATURE REVIEW

Today, one of the most important areas of use of technological innovations is the sector of health. Artificial intelligence technologies are among the frequently preferred techniques to increase efficiency in the area of health, to carry out treatment planning on time, and to diagnose diseases accurately and quickly [9]. The employing of artificial intelligence and data mining methods for the automatic identification, diagnosis, and self-management of diabetes has been extensively studied in the literature. In the literature, many studies conducted on the dataset known as "Pima Indians Diabetes" [10], and these studies aim to predict the diabetes through data mining techniques. Joshi and Shetty [11] made performance comparisons on this data set using the Bayesian approach, Naive Bayes, J28, Random forest, Random tree, REP, KNN, CART and associative rule learning algorithms. On the Pima diabetes dataset, Kumar et al. [12] utilised the Deep Neural Network technique, an unsupervised learning method, for effective diagnosis. They also used a feature selection model packaged with Extra Trees and Random Forest for selecting important features. The model performed well in comparison to other recent techniques, with an accuracy of 98.16%.

Mujumdar and Vaidehi [13] examined several machine learning methods for diagnosing diabetes using the PIMA diabetes dataset. When compared to the other machine learning methods that were being employed, linear discriminant analysis had highest accuracy of 77%. In a comparative study, Cihan et al. [14] used K-Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, Artificial Neural Network, Random Forest, and Decision Tree models. The 10-fold cross validation approach was used to evaluate the models. It was found that the Logistic Regression model produced the best results. As the performance criteria of the model, precision, sensitivity, ROC and PRC values were measured. These values are 0.76, 0.77, 0.83 and 0.83, respectively.

Another study using the same dataset is Chang et al. [15]. They used J.48, Naive Bayes and Random Forest algorithms in their diabetes classification study based on machine learning algorithms. In this study, researchers examined the results of the models without feature selection, depending on three-factor and five-factor feature selection, for examining the effect of feature selection on classification models. Among the algorithms used in the study, they observed that the Random Forest algorithm gave better results than the other two algorithms and models subjected to feature selection, with an accuracy rate of 79.57% when it was not subjected to feature selection.

For the purpose of diagnosing diabetes, Farajollahi et al. [16] compared the effectiveness of the decision tree, XGBoost, random forest, logistic regression, AdaBoost, and SVM. They clarified that among the other models, AdaBoost has the highest accuracy (83%). Sneha and Gangil [17] focused on the features selection that are involved in the early prediction of diabetes. They aimed to identify the important features and find the most appropriate machine learning classifier that provides the closest result to clinical results. According to this study, Decision tree and Random Forest had the best specificity with 98.20% and 98.00%, respectively. Naive Bayes offered the best accuracy with 82.30%.

Kumar et al. [18] developed Deep Neural Network classifier for classification. They stated that the model used gave better results than the studies in the literature with an accuracy rate of 98.16%, but the computation time was the main limitation of the study, therefore, in future studies, optimization studies for the computation time and studies on improving the computation time would make the study more effective. Ahmed et al. [19] used SVM and ANN to predict the diagnosis of diabetes. In comparison to previous published papers, their model accuracy was greater at 94.87%. Another study examining classification algorithms for diabetes using the "Pima Indians Diabetes" dataset was conducted by Karegowda et al. [20]. In this work, a hybrid model was built by combining decision tree C4.5 and k-means clustering techniques. The correct classification rate of the hybrid model run in two stages was found to be higher than the classification rate obtained using only the decision tree C4.5 method.

As can be observed, there are several studies on machine learning-based diabetes prediction in the literature. This article discusses the efficiency of the classifier algorithms for the early diagnosis of Pima diabetes dataset.

III. METHOD

A. DATA

This study used the Indian Pima dataset to predict diabetes. It is available on the general public at the UCI ML Repository. The PIMA dataset originally became available by the National Institute of Diabetes and Digestive and Kidney Diseases. The patients in this dataset were all older than 21 and all female. The dataset, which has 768 cases and 8 variables, includes information on age, blood pressure, pregnancies, skin thickness, glucose, insulin, and BMI [21]. Features of the Pima dataset are shown in Figure 1.

Abbreviation	Factor	Detail
Pr	Pregnancies	Number of times pregnant
Gl	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Bp	Blood Pressure	Diastolic blood pressure
St	Skin Thickness	Triceps skinfold thickness
In	Insulin	2-Hour serum insulin
Bm	BMI	Body mass index
Dpf	Diabetes Pedigree Function	Diabetes pedigree function
Ag	Age	Patient ages

Figure 1. The features of Pima dataset

B. INITIALIZING AND PREPROCESSING DATA

The pima dataset has passed through several steps of exploratory, analysis, and preprocessing before feeding it to the training and testing phase for diagnosing diabetes. These steps can be summarized as follows:

1. Dataset Overview

This step provides statistical description and data type for every feature of dataset. Figure 2 shows the statistical description with data type of dataset features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 2. Statistical Information and data type of Pima dataset features

2. Exploratory Data Analysis

Analysis of the dataset was done in regard to the two variables of gender and age. The patients were all female and over the age of 21, as was already mentioned. The frequencies of ages in the dataset are illustrated in Figure 3.

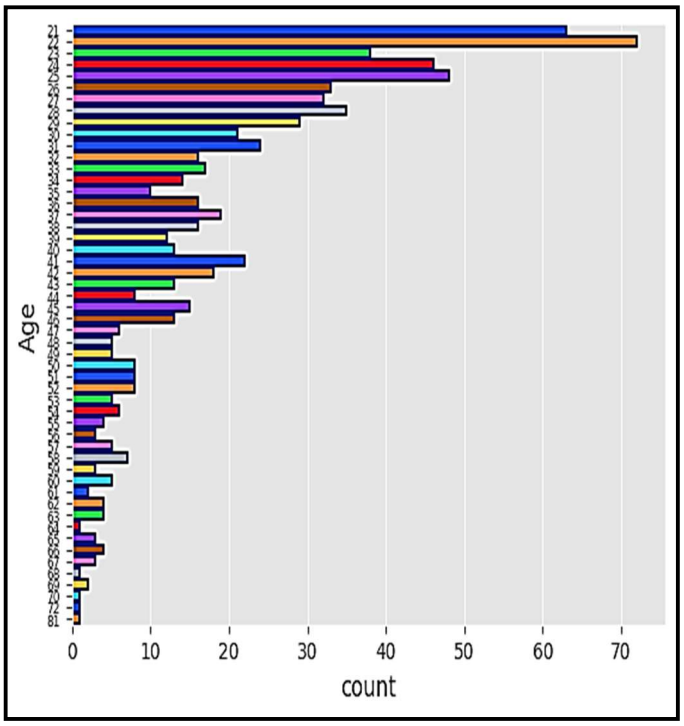

Figure 3. The frequencies of ages into dataset

3. Feature Exploration - Statistic Approach

The outliers and missing values were identified in this step. There are no missing values in the dataset. The elimination resolved the issue with outliers in each feature. The missing values and outliers in the Pima dataset are illustrated in Figures 4 and 5.

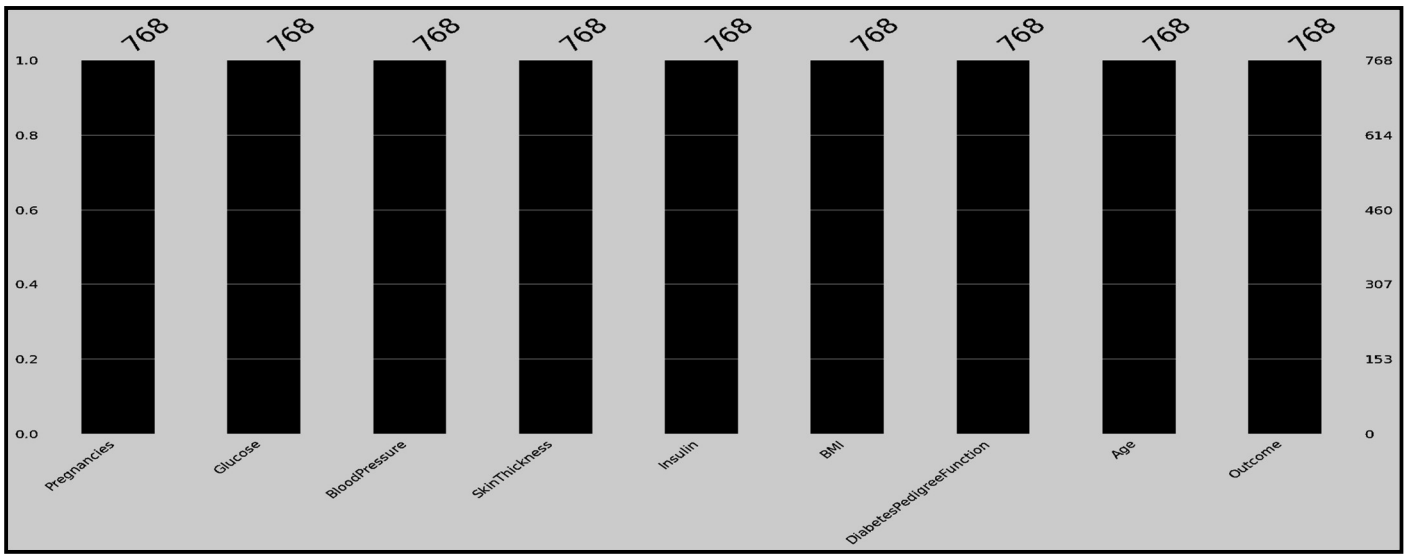

Figure 4. The missing in Pima dataset

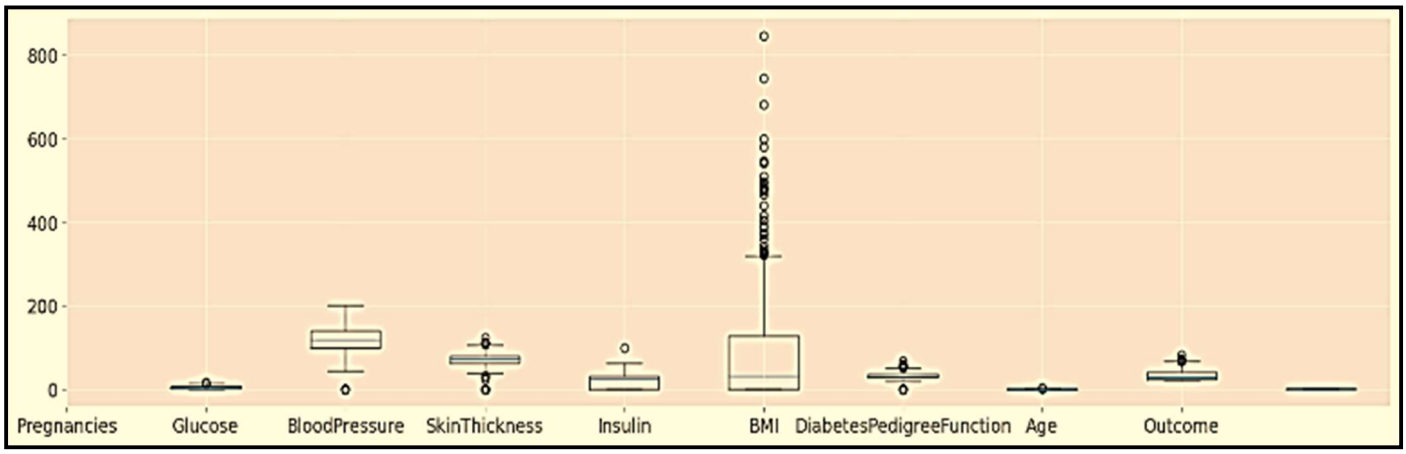

Figure 5. The outliers values in Pima dataset

4. Feature Selection

The optimal subset to represent the original dataset is chosen through the procedure of feature selection. It selects the top k features from a total of n features in the dataset by evaluating each feature with regard to the method being used [22]. By selecting the most significant and valuable attributes for the relevant problem, feature selection aims to reduce the dimensionality in the dataset.

The features that are most important are chosen in this study using the sequential backward selection (SBS) technique. The sequential backward selection (SBS) algorithm was first proposed by Marill and Green (1963). Unlike the sequential forward selection method, this algorithm operates in the reverse direction. Initially considering the entire feature set, a feature is dropped from the set at each step. in a way that optimizes the criterion function value of the current feature

subset. The removal process is repeated until the desired feature size is reached. The method in which n features are eliminated instead of one at each step is called Generalized Backward Selection. During the selection process, once the feature(s) are removed from the set, they cannot be included again. This causes the methods to give sub-optimal results [23].

The most important features obtained using SBS algorithm were ['Glucose', 'BMI', 'Age', 'Pregnancies'].

5. Imbalance Data Handling

A class is considered to be the majority class if it contains more observations in a dataset than the other class. In other words, if the observations in a database for a given class is less than that of the other class in the same database, the class is considered to be a minority class. Such datasets are called imbalanced datasets [24]. imbalanced datasets can be encountered in a wide range of practical applications, including medical diagnosis. In this work, the oversampling method using SMOTE technique used to address the issue of imbalanced data in Pima dataset.

Oversampling aims to equalize the class distribution by multiplying minority class data, as shown in Figure 6. Random oversampling is done by randomly multiplying the minority data and adding it to the original dataset. This method is simple, but it has been suggested that exact duplicates can lead to overfitting [25].

The most commonly used oversampling method is the SMOTE (Synthetic Minority Oversampling Technique) approach [26]. Unlike random sampling, this method creates synthetic data by analyzing existing minority data. SMOTE can't perfectly represent how the original samples were distributed. Byh7in the new synthetic samples. Therefore, the performance of the classifier may be impacted by the errors in the distribution of data when utilizing SMOTE-based oversampling techniques. This will increase the probability that the samples will be misclassified [27]. Figure 7. Shows the distribution of Pima dataset classes in original dataset, under-sampling, and over-sampling.

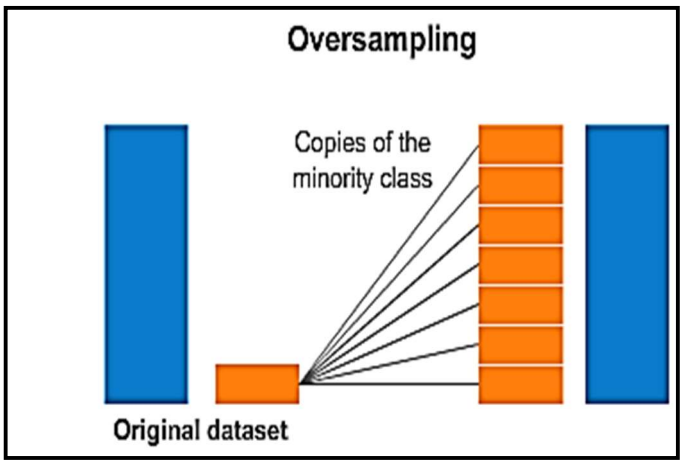

Figure 6. Oversampling dataset

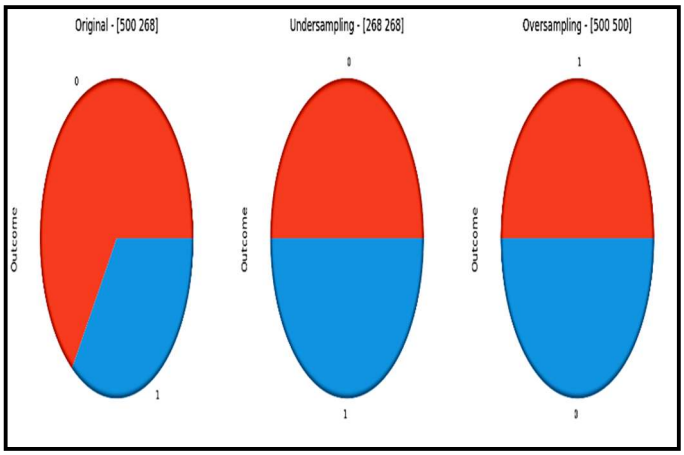

Figure 7. Under-sampling, and Over-sampling of Pima dataset

C. CONSTRUCTION MODELS

The dataset is submitted to the machine learning algorithms to classify patients after completing the data initializing and preprocessing described in the previous section. Three methods were used in this study to complete this task.

1. Multinomial Gaussian Naïve Bays

The NB method is a common probabilistic algorithm which estimates a set of probabilities by calculating the frequency and combinations of values in a dataset. The method simply applies Bayes' theorem and makes the assumption that every variable is independent of a specific class variable value. Although the method often learns rapidly in a various controlled classification tasks, the conditional independence assumption is rarely true in practical implementations [28]. The mathematical expression for Bayes' theorem is given in Eq. 1.

$$P(H\backslash D)= (P(H)P(D\backslash H))/(P(D)), \quad (1)$$

where, the probability of occurrence of event H when the probability of occurrence of the event D is known is P(H\D). The significant benefit of NB is that it needs not much measure of training data. In supervised machine learning, the mathematical expression is represented in Eq. 2.

$$P(H\backslash D)=P(x_i,\ldots\ldots,x_n \mid H)=\prod_i P(x_i \mid H), \quad (2)$$

where, x_i ,… , x_n represent the input attributes that the conditional probabilities compute according to the known probabilities of the target variables in the training dataset.

One popular approach for predicting diabetes is the Multinomial Naive Bayes classifier. It is built on the concepts of Bayes' theorem and supposes that given the class (diabetes or not diabetes), the features in the diabetes dataset are conditionally independent. It is computationally effective and frequently works well in practice in spite of this simplifying assumption. The Multinomial Naive Bayes classifier is used when the data follows a multinomial distribution [29].

2. Logistic Regression

Logistic regression analysis is one of the methods applied to assign observations to groups in the data set. In statistics, logistic regression is a technique utilized to classify two-class variables [30]. The relationship among a set of independent variables and a categorical dependent variable is represented by a curve in logistic regression, which measures the probability that a given event will occur [31]. While independent variables may have continuous or categorical values, the dependent variable must be categorical. The use of logistic regression is appropriate for models that concentrate on binary results. such as success or failure, yes or no, healthy or unhealthy, etc. in the decision-making situation of an event, rather than the time of occurrence of the events. Logistic regression assigns these categorical values as 1 if "Yes" and 0 if "No" [32]. The sigmoid function is typically used in logistic regression to obtain binary output probabilities based on one or more factors and to choose the most suitable parameters. The sigmoid function (σ) and the sigmoid function input (z) are shown in Eq. 3 below [33].

$$\sigma(z)=1/(1+e^{(-z)}), \quad (3)$$

where, zϵR and $\boldsymbol{\sigma}(z)\epsilon(0,1)$

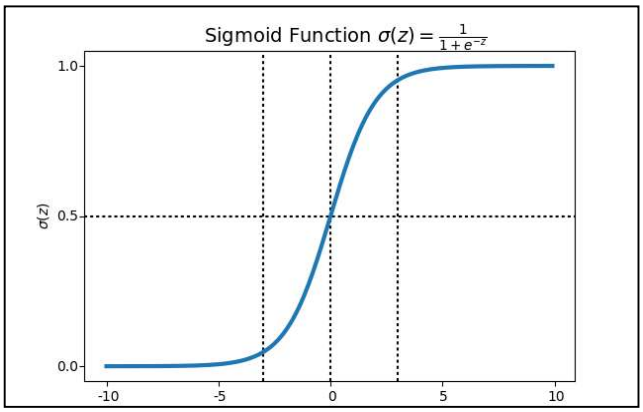

Figure 8. Sigmoid Function

3. Multi-Layer Perceptron Algorithm

Multi-Layer Perceptron MLP is a feedforward neural network that contains one or more layers between the input and output layers. Every neuron in the layers is connected to each neuron in the adjacent layers. The structure of an artificial neuron is shown in Figure 9. The neuron calculates the weighted sum of n inputs, adds a threshold value, and then uses an activation function to the result to calculate the output [34, 35].

$$S=\sum x_i w_i + w_0, \quad (4)$$

$$Y=f(s), \quad (5)$$

The activation function known as the sigmoid is the most popular and is defined by Eq. 3. The effectiveness of the neural network model depends on the nonlinearity of this function [36]. Additionally, the function scales the output to the range [0-1].

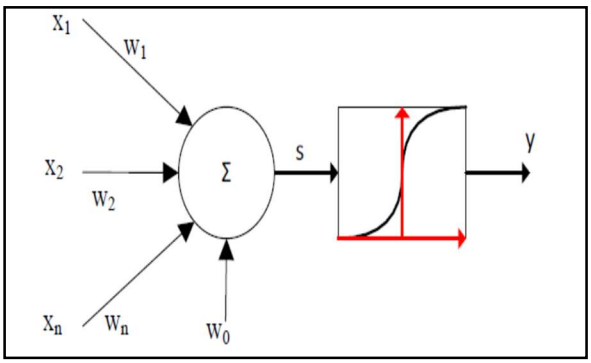

Figure 9. Artificial neuron

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The classification performances of all models are discussed in this section. The classification performance of the models was evaluated using the metrics accuracy, precision, recall, and f1-score. According to the confusion matrix values (Fig.10) obtained from each algorithm, these metrics [8, 36] are given in equations. 6 to 9 [37].

$$Accuracy= (TP+TN)/(TP+TN+FP+FN), \quad (6)$$

$$Recall= TP/(TP+FN), \quad (7)$$

$$Precision= TP/(TP+FP), \quad (8)$$

$$F\_measure = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}. \quad (9)$$

where, TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative. The results obtained from the machine learning algorithms with Pima dataset are given in Tables 1 to 6.

Table 1. Evaluation metrics of Multinomial Naive Bayes algorithm on training dataset

	0	1	accuracy	macro avg.	weighted avg
'Precision'	0.605	0.664	0.626	0.635	0.634
'Recall'	0.761	0.487	0.626	0.624	0.626
'f1-score'	0.674	0.562	0.626	0.618	0.619
'Support'	406	394	0.62625	800	800

Table 2. Evaluation metrics of Multinomial Naive Bayes algorithm on testing dataset

	0	1	accuracy	macro avg.	weighted avg.
precision	0.567	0.699	0.615	0.632	0.639
recall	0.766	0.481	0.615	0.624	0.615
f1-score	0.652	0.569	0.615	0.611	0.608
support	94	106	0.615	200	200

Table 3. Evaluation metrics of Logistic Regression algorithm on training dataset

	0	1	accuracy	macro avg.	weighted avg.
precision	0.739	0.751	0.745	0.745	0.745
recall	0.768	0.721	0.745	0.745	0.745
f1-score	0.754	0.736	0.745	0.745	0.745
support	406	394	0.745	800	800

Table 4. Evaluation metrics of Logistic Regression algorithm on testing dataset

	0	1	accuracy	macro avg.	weighted avg.
precision	0.760	0.820	0.79	0.790	0.792
recall	0.809	0.774	0.79	0.791	0.790
f1-score	0.783	0.796	0.79	0.790	0.790
support	94	106	0.79	200	200

Table 5. Evaluation metrics of Multi-Layer Perceptron algorithm on training dataset

	0	1	accuracy	macro avg.	weighted avg.
precision	0.793	0.774	0.783	0.784	0.784
recall	0.776	0.792	0.784	0.784	0.784
f1-score	0.785	0.783	0.784	0.784	0.784
support	406	394	0.784	800	800

Table 6. Evaluation metrics of Multi-Layer Perceptron algorithm on testing dataset

	0	1	accuracy	macro avg.	weighted avg.
precision	0.780	0.789	0.785	0.784	0.785
recall	0.756	0.811	0.785	0.783	0.785
f1-score	0.768	0.800	0.785	0.784	0.785
support	94	106	0.785	2000	200

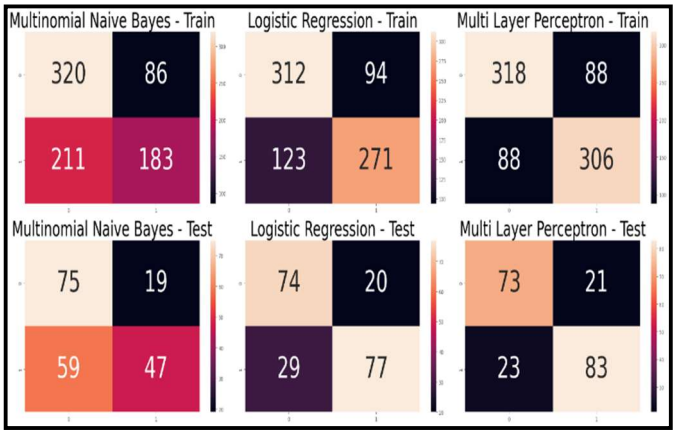

Figure 10. Confusion matrix of machine learning algorithms

The tables above present the classification results for the methods Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Multi-Layer Perceptron (MLP) achieved by dividing the dataset into 80% training dataset and 20% testing dataset. The algorithm MLP, which achieved accuracy of 78% on the training dataset and 78.5% on the testing dataset, offered the best classification performance. The algorithm logistic regression LR, which has accuracy of 74.5% on training datasets and 79% on testing datasets, has the second-highest performance in classification. The technique of multinomial naive Bays (MNB) placed third with performance accuracy of 63% on the training dataset and 62% on the testing dataset.

V. CONCLUSIONS

Early diagnosis of diabetes is essential for managing a healthy life as the condition is an ongoing disease characterized by unusually high blood glucose levels. In this study, the performances of classifiers Multinomial Naive Bayes, Logistic Regression, and Multi-Layer Perceptron for early diagnosis of this disease are discussed. As a result, the MLP technique performed better than other classifiers, with 78% classification accuracy. This classifier misclassifies 88 out of 394 positive samples and 88 out 406 negative samples in the training dataset in experiments that divided the Pima dataset into 80% training dataset and 20% testing dataset. While in the test dataset, this classifier incorrectly labels 21 out of 94 negative samples and 23 out of 106 positive samples. The differences in results of studies that used the same dataset can be related to different dataset preprocessing procedures or different hyper parameter tuning of the models. The dataset analysis and preprocessing in this study helped to clean the dataset and produce the best results. The future studies might focus on selecting the most important features of the disease and conducting experiments based on optimization algorithms.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] A. M. Egan & S. F. Dinneen, "What is diabetes?" *Medicine (United Kingdom)*, vol. 42, issue 12, pp. 679–681, 2014. https://doi.org/10.1016/j.mpmed.2014.09.005.

[2] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021. https://doi.org/10.1016/j.ijcce.2021.01.001.

[3] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia Computer Science*, vol. 112, pp. 2519–2528, 2017. https://doi.org/10.1016/j.procs.2017.08.193.

[4] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O.B. Pineda, "Diabetes diagnostic prediction using vector support machines," *Procedia Computer Science*, vol. 170, pp. 376–381, 2020. https://doi.org/10.1016/j.procs.2020.03.065.

[5] J. Chaki, S. T. Ganesh, S. K. Cidham, & S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, issue 6, pp. 3204-3225, 2022. https://doi.org/10.1016/j.jksuci.2020.06.013.

[6] T. Sharma, and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, issue 1, p. 30, 2021. https://doi.org/10.1186/s42492-021-00097-7.

[7] S. Afzali and O. Yildiz, "An effective sample preparation method for diabetes prediction," *International Arab Journal of Information Technology*, vol. 15, no. 6, 2018.

[8] E. M. Hameed, I. S. Hussein, H. G. Altameemi, & Q. K. Kadhim, "Liver disease detection and prediction using SVM techniques," *Proceedings of the 2022 3rd IEEE Information Technology to Enhance e-learning and Other Application (IT-ELA)*, 2022, pp. 61-66. https://doi.org/10.1109/IT-ELA57378.2022.10107961.

[9] F. Al-Areqi and M. Z. Konyar, "Effectiveness evaluation of different feature extraction methods for classification of Covid-19 from computed tomography images: A high accuracy classification study," *Biomedical Signal Processing and Control*, vol. 76, 2022, https://doi.org/10.1016/j.bspc.2022.103662.

[10] Kaggle database. [Online]. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[11] S. Joshi, S. R. Priyanka Shetty, "Performance analysis of different classification methods in data mining for diabetes dataset using WEKA tool," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, issue 3, pp. 1168-1173, 2015. https://doi.org/10.17762/ijritcc2321-8169.150361.

[12] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clinical Epidemiology and Global Health*, vol. 7, issue 4, pp. 530–535, 2019. https://doi.org/10.1016/j.cegh.2018.12.004.

[13] A. Mujumdar, V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Proc Comput Sci.*, vol. 165, pp. 292-299, 2019. https://doi.org/10.1016/j.procs.2020.01.047.

[14] P. Cihan and H. Coskun, "Performance comparison of machine learning models for diabetes prediction," *Proceedings of the 29th Signal Processing and Communications Applications Conference (SIU'2021)*, Istanbul, Turkey, 2021, pp. 26–30. https://doi.org/10.1109/SIU53274.2021.9477824.

[15] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput & Applic.*, vol. 35, pp. 16157–16173, 2023. https://doi.org/10.1007/s00521-022-07049-z.

[16] B. Farajollahi, M. Mehmannavaz, H. Mehrjoo, F. Moghbeli, M. J. Sayadi, "Diabetes diagnosis using machine learning," *Front Health Inform*, 2021. https://doi.org/10.30699/fhi.v10i1.267.

[17] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, article 13, pp. 1–19, 2019. https://doi.org/10.1186/s40537-019-0175-6.

[18] P. B. M. Kumar, R. S. Perumal, R. K. Nadesh, and K. Arivuselvan, "Type 2: Diabetes Mellitus prediction using deep neural networks classifier," *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 55–61, 2020, https://doi.org/10.1016/j.ijcce.2020.10.002.

[19] U. Ahmed et al., "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529-8538, 2022. https://doi.org/10.1109/ACCESS.2022.3142097.

[20] A. G. Karegowda, V. Punya, M. A. Jayaram, A. S. Manjunath, "Rule based classification for diabetic patients using cascaded k-means and decision tree C4.5," *International Journal of Computer Applications*, vol. 45, issue 12, pp. 45-50, 2012.

[21] E. M. Hameed, & H. Joshi, "Current diabetes classification and prediction models using intelligent techniques," Proceedings of the VI. International Scientific Congress of Pure, Applied and Technological Sciences, MINAR CONGRESS 6, 2022, pp. 20-50. https://doi.org/10.47832/MinarCongress6-2.

[22] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[23] J. Kittler, "Feature set search algorithms," In: C.H. Chert, Ed., Pattern Recognition and Signal Processing, Sijthoff and Noordhoff, Mphen aan den Rijn, Netherlands, 1978, pp. 41–60. https://doi.org/10.1007/978-94-009-9941-1_3..

[24] J. Gong, & H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Computational Statistics & Data Analysis*, vol. 111, pp. 1-13, 2017. https://doi.org/10.1016/j.csda.2017.01.005.

[25] G. E. Batista, R. C. Prati, & M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, issue 1, pp. 20-29, 2004. https://doi.org/10.1145/1007730.1007735.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. https://doi.org/10.1613/jair.953.

[27] Z. Zheng, Y. Cai, & Y. Li, "Oversampling method for imbalanced classification," *Computing and Informatics*, vol. 34, issue 5, pp. 1017–1037, 2016. [Online]. Available at: https://www.cai.sk/ojs/index.php/cai/article/view/1277.

[28] G. Dimitoglou, J. A. Adams, & C. M. Jim, "Comparison of the C4.5 and a Naive Bayes classifier for the prediction of lung cancer survivability index terms-data mining, mining methods and algorithms, text mining," *Journal of Computing*, vol. 4, issue 8, pp. 1-9, 2012. https://doi.org/10.48550/arXiv.1206.1121.

[29] M. A. Uddin, M. M. Islam, M. A. Talukder, M. A. A. Hossain, A. Akhter, S. Aryal, & M. Muntaha, "Machine learning based diabetes detection model for false negative reduction," *Biomedical Materials & Devices*, pp. 1-17, 2023. https://doi.org/10.1007/s44174-023-00104-w.

[30] S. Kost, O. Rheinbach, and H. Schaeben, "Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling," *Geochemistry*, no. September, p. 125826, 2021, https://doi.org/10.1016/j.chemer.2021.125826.

[31] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nurs.*, vol. 43, no. 2, pp. 154–164, 2013, https://doi.org/10.4040/jkan.2013.43.2.154.

[32] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *J. Data Anal. Inf. Process.*, vol. 7, no. 4, pp. 190–207, 2019, https://doi.org/10.4236/jdaip.2019.74012.

[33] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," *Proceedings of the IEEE Int. Conf. Comput. Netw. Informatics, ICCNI'2017*, vol. 2017 – January, 2017, pp. 1–9, https://doi.org/10.1109/ICCNI.2017.8123782.

[34] Y. Kumar, G. Sahoo, "Analysis of Bayes, neural network and tree classifier of classification technique in data mining using WEKA," *Proceedings of the Second International Conference on Computer Science & Information Technology (CS & IT)*, 2012, pp. 359-369. https://doi.org/10.5121/csit.2012.2236.

[35] D. Morariu, R. Creţulescu, M. Breazu, "The weka multilayer perceptron classifier," *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*, vol. 7, issue 1, 2017.

[36] E. M. Hameed, H. Joshi, and Q. K. Kadhim, "Advancements in artificial intelligence techniques for diabetes prediction: A comprehensive literature review," *Journal of Robotics and Control (JRC)*, vol. 6, issue 1, pp. 345-365, 2025. https://doi.org/10.18196/jrc.v6i1.22258.

[37] E. M. Hameed, and H. Joshi, "Performance comparison of machine learning techniques in prediction of diabetes risk," *AIP Conference Proceedings*, vol. 3051, no. 1, AIP Publishing, 2024. https://doi.org/10.1063/5.0191611.

***EMAD MAJEED HAMEED,** Master of Computer Science / Informatics and currently Ph.D. scholar of computer science in Gujarat university, India. He works as instructor in Middle technical university, Iraq. Research Interests: machine learning, pattern recognition, image processing.*

He can be contacted at email: emadhameed@gujaratuniversity.ac.in, emadmajeed@mtu.edu.iq

***HARDIK JOSHI** is Dr. Asst. Professor with the Department of Computer Sc., Gujarat University, India. He teaches and supervises students of Ph.D., MCA & M.Tech of Gujarat University. His research area includes Natural Language Processing & Information Retrieval. He can be contacted at email: hardikjoshi@gujaratuniversity.ac.in*

***Hayder Jasim Habil,** Director of the Security Permits Department at the Middle Technical University Teaching at the College of Applied Arts*

1. Technical Diploma, Technical Instructors Training Institute, Electronics Department.

2. Bachelor's Degree, Faculty of Electrical Technology, Computer Technology Department.

3. Master's degree in computer technology engineering, Al-Razi University, Islamic Republic of Iran.

*He can be contacted at email: **hayder-jasim@mtu.edu.iq***

Dr. Mudhar A. Al-Obaidi is a lecturer in computing at the Middle Technical University, Iraq. He obtained his BSc and MSc degrees in chemical engineering from the University of Baghdad and University of Technology in Iraq in 1993 and 1997, respectively. He obtained his PhD in chemical engineering in 2018 from the University of Bradford, UK.

*He has contributed to more than 65 peer-reviewed journal papers, conference presentations, and chapters. Recently, he has published his first book related to wastewater treatment. He can be contacted at email: **dr.mudhar.alaubedy@mtu.edu.iq***