Alkadhim Journal for Computer Science, Vol. 2, No. 4 (2024)

Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

IRAQI Academic Scientific Journals

Alkadhim Journal for Computer Sciences (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

KJCS ALKADHIM JOURNAL FOR COMPUTER SCIENCE

Multiclass Diabetes Prediction Using Ensemble learning Techniques

Akbal O. Salman

Space Technology Engineering Department, Electrical Engineering Technical College, Middle Technical University, Iraq.

Article information

Article history:

Received: Nov ,7,2024

Accepted: Dec, 8,2024

Available online: Dec, 25, 2024

Keywords:

Diabetes,
PIMA dataset,
Naïve Bayes,
Random Forest,
SVM,
Multiclass diabetes dataset,

Corresponding Author:

Akbal O. Salman

akbal.o.salman@mtu.edu.iq

DOI:

https://doi.org/10.61710/kjcs.v2i4.87

Abstract

Diabetes is one of the most prevalent diseases in the modern era, leading to a significant number of deaths annually, as reported by the World Health Organization. Early prediction of diabetes can substantially improve patient outcomes and save lives. This study introduces a new model for predicting diabetes using the Random Forest algorithm, known for its powerful ability to split data until reaching an optimal state. Two datasets are utilized: the Multiclass Diabetes dataset and the PIMA Indian Diabetes dataset. The data are preprocessed by removing outliers, handling missing values, and balancing the classes. These preprocessed data are then classified using the Random Forest algorithm through continuous splitting until the stopping criteria are met, aiming to predict diabetic individuals. The proposed model demonstrated superior performance with the Multiclass Diabetes dataset, it achieves a validation accuracy of 100%, a precision of 98.20%, and recall and F1 scores of 98.11% and 98.12%, respectively. With the PIMA dataset, the proposed model achieves a validation accuracy of 85.30%, with precision, recall, and F1 scores of 88.07%, 87.50%, and 87.50%, respectively. In addition to our proposed model, we built many machine learning models with the first dataset such as SVM, logistic regression, logistic regression with L1/ L2 regularization, K-NN, and naïve bayes. Our results indicate that the Random Forest algorithm significantly outperforms other machine learning techniques in predicting diabetes, offering a highly accurate and reliable tool for early diagnosis. This research underscores the potential of ensemble learning in healthcare, particularly in managing chronic diseases like diabetes.

1. Introduction

Introduction Diabetes is one of the most chronic diseases of our time and is associated with increased blood sugar levels. According to reports and statistics from the Public Health Organization, the number of people with diabetes is estimated at approximately 422 million people, while the number of deaths due to this disease exceeds one and a half million deaths annually [1]. The early detection of diabetes can help to avoid the degradation of their health cases, although the diabetes has no completely cure yet. Therefore, proposing a system that has the ability to detect the diabetes cases in three types based on machine learning can be very helpful to people with diabetes or those at risk of infection. Previously, many studies have been conducted to predict the infect of diabetes using machine learning [2] and deep learning [3][4]. In the case of deep learning, convolutional neural network (CNN) [5] in addition to the variants of CNN [6] such as VGG, Inception, Exception, Inception-ResNet-V2, and etc. also used for this task. In the case of machine learning, ensemble

learning is considered one of the most effective machine learning techniques. In this technique, multiple models (also called weak or base learners) are incorporated after being trained to produce a new strong and precise model named ensemble by leveraging
the merits and viewpoints of each one of these incorporated models. Ensemble learning assists in raising the performance model and making it more generalized through reducing over fitting [7], [8]. Random forest is one of the techniques that used ensemble learning [9].
Moreover, integrating Internet of Things (IoT) technology into healthcare systems has opened new horizons in diabetes management. IoT devices, such as wearable sensors and smart glucometers, can continuously monitor blood sugar levels, physical activity, and other vital signs, providing real-time data to healthcare providers and patients [10][11]. Beside ensemble random forest, many other machine learning algorithms have been employed for predicting diabetes, such as support vector machine (SVM) [12], K- nearest neighbor (K-NN) [13], logistic regression [14].
In this study, a proposed model based on the random forest technique to predict diabetes based on two datasets; Multiclass Diabetes Dataset [15] and PIMA Indian Diabetes dataset [16], where some challenges are faced with this work like the dataset is class imbalanced, additionally, the Multiclass Diabetes Dataset is very recent and this is the first study that used this dataset, thus, no previous studies available for working on this dataset. The discussion is based on analysis the data through using visualization tools, where principal component analysis (PCA) [17] is employed to plot the target classes of the dataset. Various performance evaluation metrics have been used to assess the performance of the proposed model, including accuracy, F1 score, precision, and recall. The proposed model achieves better performance compared to previous studies conducted on the same dataset. The Random Forest algorithm was chosen as the primary algorithm for this study due to its ability to perform effective splitting and select the best solutions. It is recognized as one of the best algorithms in the ensemble learning domain.
The contribution of this study is illustrated as follows:

- Applied multiple machine learning techniques for multiclass diabetes prediction on a multiclass diabetes dataset, which is being used for the first time.
- Two datasets were used for testing our proposed model: binary and multiclass datasets.
- Multiple preprocessing operations have been used including handling missing values, removing outliers, and class balancing.

2. Related Works

Many studies had been inducted to diagnose diabetes, some of most important studies are shown below:

in [18], proposed many machine learning models for classifying a diabetes or non-diabetes. They used decision tree, SVM, Random Forest, and K-NN classifiers to classify a collected dataset which composed of more the 20,000 records, with 10 attributes. Accuracy, f1 score, precision, and recall are the evaluation metrics that are used to measure the performance of those models. The best results they gained were 97.5% for accuracy, 97% for f1 score, 97.4 and 96.6 for precision and recall, respectively. O. I. In [19] detected and classifying the diabetes by using five machine learning models; K-NN, logistic regression, Bernoulli Naïve Bayes, SVM, and decision tree classifiers. They used PIMA dataset to achieve their work. The performance of these models gave the following results: K-NN gave the best performance with accuracy 79.6%, BNB 77.2%, Logistic Regression 72.7%, SVM 71.7%, and finally decision tree 63%. In [20] explored the most well-known methods such as SVM, DNN, etc. that were used to discover diabetes. PIMA dataset was used for checking and examining these methods. With each model, they altered the parameters and preprocessed the data in the dataset. After all the processing operations, they found that the best accuracy performance gained is 77.86% using 10-fold cross-validation. In [21] proposed a classification model for detecting diabetes in its early stages. Moreover, they used an IoT-based model to check and monitor the diabetes disease. The proposed methods that were used to achieve this goal are logistic regression, random forest, and multilayer perception. PIMA dataset was used with this model. The accuracy, f1 score, precision, and recall were used for measuring the performance of this mode. The random forest achieved 77.4% accuracy, while the other methods exceeded this rate. In [22] built a decision tree classification system. They used an open dataset with 768 cases. At first, they divided the dataset into training and testing categories, with a proportion of 70% of data for training, and 30% for testing. Lastly, the proportion was modified to 50% for testing and 50% for training. They used the accuracy metric for measuring the performance of the model, and the results were 63% for the first proportion, and 71.3% for the last proportion. Below we'll illustrate the weaknesses and gapes that are revealed in the previous studies: In [18], the authors

used a real dataset obtained directly from the source, which is difficult to access. Additionally, the dataset was split into only two parts: training and testing, with no validation part. Consequently, the model was not examined on unseen data. In [19], the PIMA dataset was used. The dataset was divided into two sets: 80% for training and 20% for testing, without a validation set, thus it did not test on the unseen data. The model's performance was not very strong, with the best model accuracy being 79.6% achieved by K-NN. This indicates that the data needed more preprocessing. In this case, ensemble methods are recommended; therefore, it would have been better if a random forest had been used. In [20], the authors applied their system to the PIMA Indian diabetes dataset using a deep convolutional neural network (DNN) and SVM to build the proposed model. They were split the data into two sets without a validation set, and the performance was not very strong. The authors in [21] used multi-layer perceptron (MLP) and long short-term memory (LSTM) methods to predict diabetes. They applied their models to the PIMA dataset, achieving an accuracy of 86.08% and values of 86.6% for precision and 85.1% for recall for MLP, while LSTM achieved 87.26% accuracy and 2.285 RMSE. The values of precision and recall were not computed for LSTM. Given the class imbalance, accuracy is not the optimal metric to measure performance; precision, recall, and F1 score are better metrics. The authors used only precision and recall for evaluation, obtaining values of 86.6% for precision and 85.1% for recall with MLP. For the random forest, they obtained an accuracy of 77.4%, with values of 76.9% for precision and 75.7% for recall.

In [22], the PIMA dataset was used for training the proposed model. However, their model has many gaps. They split the dataset twice; in the first split, they divided the dataset into two parts: 70% for training and 30% for testing. In the second split, they divided the dataset fifty-fifty for training and testing sets, although, the did not justify the reason behind these two splits. This approach is not effective and causes high over fitting. They did not perform clear preprocessing on the dataset, as there was no manipulation of missing values or outliers. Consequently, the performance of their model is weak. The following steps have been conducted to address these gaps:

- This study focuses on testing the validation set, which contains unseen data, to avoid bias towards the tested data during the training-testing process.
- Multiple machine learning techniques were used, including the ensemble learning technique (Random Forest), which continues trying until it reaches the optimal solution.
- Two datasets were used to prove the strength of this model, and one of them is a multiclass dataset.

3. Theoretical Background

This section presents simple background about the techniques that have been used in this study:

1) Random Forest (RF): RF is a classification method that employs an ensemble learning technique to achieve its goal. RF relies on the collaboration of individual classifiers, known as “weak learners,” to create “strong learners” by constructing a large number of decision trees (DTs). These trees are decorrelated, forming what is known as a random forest. In RF, the solution to any problem is not based on a single DT. Instead, RF aggregates outputs from multiple shallow trees. This process involves a technique called bagging, which stands for bootstrap aggregating. By using bootstrapping to create multiple datasets from the original data, bagging builds n predictors. These predictors, developed from independently sampled trees, are combined through an averaging process. This process helps to solve problems whether they are related to estimation or classification by combining the predictions from all the trees. The name “forest” comes from using multiple DTs to produce the final decision for the classification task [2]. For classification, the general equation of the ensemble learning to build RF by in the case of voting is expressed in Equation 1 [23].

$$f(x) = \arg\max_{y \in \mathcal{Y}} \sum_{j=1}^{J} I(y = h_j(x)) \qquad (1)$$

Where $h_j(x)$ represent the base learners, $f(x)$ denotes the ensemble predictor, $\mathcal{Y}$ is the set of possible values, J is the total number of ensemble’s base models.

2) Logistic regression (LR): LR is a supervised machine learning method that is not related to traditional regression and is not used to solve regression problems. LR is used for predicting binary target classes by estimating how the dependent variable, as assumed by the regression function, is connected with at least one independent variable. This is done by estimating the likelihood with the assistance of the sigmoid function. Since the response variable is binary, it is not normally distributed and tends to be nonlinear. The dependent variable is

dichotomous, taking the form of (0/1, -1/1, true/false), while the independent variable may be one of the following types: binomial, ordinal, interval, or ratio-level. Equation 2 is used to calculate LR [24].

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}} \quad (2)$$

Where L represents the upper bound, while k is the growth rate of the logistic, x is the independent variable, and x_0 is the midpoint of sigmoid for the value of x.

3) Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning method that can be employed for both classification and regression tasks, though it is predominantly used for classification. SVM works by finding the optimal hyperplane in a multidimensional space that best separates different classes of data points. This hyperplane acts as a boundary for classifying the data points. The primary goal is to maximize the margin between the hyperplane and the nearest data points from each class, ensuring the most distinct separation possible [25].

4) Naïve Bayes: Naïve bayes (NB) is one of the most effective classifiers in machine learning. It is a probabilistic classification approach based on Bayes' theorem, which deals with probabilities. In Naïve Bayes, the assumption is that the features within the same class are independent of each other, meaning the presence or absence of a feature does not influence the presence or absence of another. This algorithm performs exceptionally well on datasets with missing values and imbalanced target classes [26], [27]. Equation 3 shows how to calculate the posterior priority:

$$p(c|x) = \frac{p(x|c)\, p(c)}{p(x)} \quad (3)$$

Where,
$p(c|x)$ represents the posterior probability
$p(x|c)$ is the likelihood
$p(c)$ is the class prior probability
$p(x)$ is the predictor prior probability

5) K-Nearest Neighbour: K-Nearest Neighbour (KNN) is a straightforward and efficient classification algorithm that can also be used for regression tasks. This method is known for its low time complexity. The Euclidean distance equation is employed to calculate the distance between data points, for both existing and new points. In this algorithm, after splitting the dataset into training and testing sets, it assigns a label to each object in the testing set by identifying a set of k objects in the training set that are closest to the testing object. The label assigned is based on the most predominant class among these k nearest neighbours [28].
The performance of this algorithm is significantly influenced by three factors: the value of k, the distance metric used to calculate the proximity between objects, and the method of assigning labels to the objects [29].

6) Logistic Regression with L1/ L2 Regularization: The goal of employing regularization in logistic regression is to control the complexity of the model by adding a penalty to its loss function. The two most well-known types of regularization in logistic regression are L1 and L2 regularization. These two types have different effects on the model. L1 regularization affects the model by adding a penalty to the loss function that is equivalent to the sum of the absolute values of the coefficients. While the penalty added by L2 regularization is equivalent to the sum of the squares of the coefficients. The L1 regularization has a significant impact on feature selection, as it aims to set some coefficients' values to zero. Consequently, it selects a simpler model that works with only a subset of features. Thus, most features have zero coefficients, except for a few. In the case of L2 regularization, it shrinks all large coefficients towards zero, but not exactly to zero. This leads to improved generalization, making the model less sensitive to small variations or changes in the input data. As a result, all features are retained, but those with less importance have smaller weights [30].

4. Methods and Materials

4.1 Dataset

Two datasets were used in this study, the multiclass Diabetes Dataset [15] and Pima Indians Diabetes Database [16], the former is the most recent dataset in this field, while the latter is the most well-known dataset, with many studies conducted

using it. Multiclass Diabetes Dataset is used in this study to construct a prediction model that is able to predict the cases of the people whether they are diabetic, non-diabetic, or predict-diabetic. This dataset is composed of 1000 samples, and these samples divided into 844, 103, and 53 for diabetic, non-diabetic, and predict-diabetic, respectively. The dataset contains 11 features described in Table 1.

Table 1: The description of the multiclass diabetes dataset.

Feature	Description
Gender	Detect the sex of each sample if it is male or female
Age	Detect the age of the patient
BMI	body mass index; a metric that checks whether a patient has a normal weight or not
Chol	Cholesterol
TG	Indicate the triglyceride values for each patient
Urea	Urea proportion
VLDL	Very Low-Density Lipoprotein Cholesterol
Cr	Creatinine
LDL	Low-Density Lipoprotein Cholesterol
HbA1c	Hemoglobin A1c
HDL	High-Density Lipoprotein Cholesterol; the good cholesterol
Target class	Classify the patients: diabetic, non-diabetic, or predict-diabetic

Pima Indians Diabetes Database is a dataset with binary target classes, which is composed of many features to check whether the patient have diabetes or not. The samples of PIMA are 768, each sample represents a woman. This dataset composed of eight attributes in addition to one target class. Table 2 describes the features of this dataset.

Table 2: The description of the Pima Indians Diabetes Database

Feature	Description
Pregnancy	The value in this feature refers to how many times each patient has been pregnant.
Glucose	Represent the value of blood sugar of each patient.
Blood Pressure	Indicates the blood pressure for each patient
Skin Thickness	Represents the thickness of the skin of each patient, measured in millimeters, based on their nutritional status.
Insulin	Indicates the level of insulin in the blood for each patient.
BMI	body mass index; a metric that checks whether a patient has a normal weight or not
Diabetes Pedigree Function	The probability of getting diabetes based on family history and genetic factors.
Outcome	Detect the class of each patient; diabetes or not

4.2 The Proposed Model

This model is composed of four main stages, starting with the preprocessing stage, then training and testing stage, to the prediction stage, and finally, performance evaluation stage. The RF technique is the classifier that is used to make the prediction, at the first, the dataset is preprocessed by using multiple steps, which are handling of both missing and outliers' values, and since the target classes are not balanced, oversample is added to make the target classes balanced. Figure 1 shows the proposed model.

The dataset is split to three groups; the training set which represents 70% of the size of the dataset, testing data which represents 20% of the dataset, and finally, validation data which represents 10% of the dataset. After all these processes, the training and testing processes are starting, where the system decides how many trees are required to achieve the goal, and the features that are required in the training and testing data. Boosting sampling is the next step, in which random samples

are chosen from the training set to control the splitting process of the trees based on the features that are selected at each split. The trees are built based on the bootstrap and the selected features. The training process is continued for every tree by choosing randomly subset of tree and continue splitting until meet the stopping criteria. Bagging ensemble is used, where the sub-trees are created and based on multiple subsets of training data and make the decision through taking the majority vote among all these models that have been created, the latest steps represent prediction. At the end, the outcomes are validated using evaluation metrics accuracy, f1 score, precision and recall using evaluation metrics.

In addition to our proposed model, many models have been built and implemented with the same dataset such as SVM, logistic regression, logistic regression with L1/ L2 regularization, K-NN, and naïve bayes. Finally, to compare this study with the previous works, we also implemented our proposed model with the PIMA dataset.

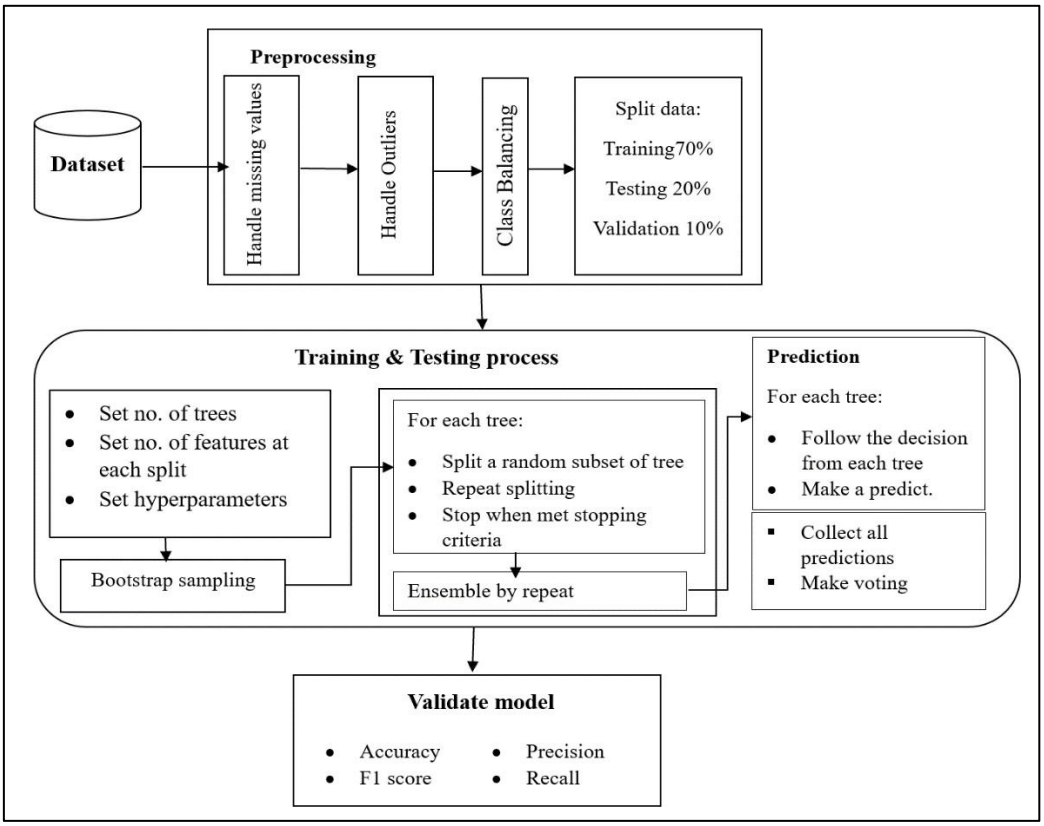


Figure 1: The proposed ensemble model

5. Experimental Results and Discussion

This section provides a detailed explanation of the experimental results of this study, including the measurements used, the results achieved, and a discussion of the outcomes obtained from the implementation of this study.

5.1 Evaluation Metrics

The performance of the methods used in this study is evaluated using the following metrics: accuracy, precision, recall, and F1 score [31]. Accuracy measures the proportion of correctly classified samples out of the total samples in the dataset. However, it is a reliable metric only when the dataset is balanced. The calculation of accuracy is shown in Equation 4.

$$= \frac{+}{+ + +} \quad (4)$$

Where:

represent true positives and true negatives, respectively, indicating correct predictions, in which is positive prediction for diabetes, while the healthy cases indicated by . Conversely, stand for false positives and false negatives, respectively, representing incorrect predictions.

Precision calculates the ratio of correctly predicted positive labels to all labels predicted as positive. It is determined using Equation 5.

$$= \frac{}{+} \quad (5)$$

Recall measures the model's ability to correctly identify positive samples. It indicates how well the model captures all the true positive cases. The calculation of recall is shown in Equation 6.

$$= \frac{}{+} \quad (6)$$

F1 score measures the harmonic mean of precision and recall, giving equal weight to both. A higher F1 score indicates better model performance. The calculation of the F1 score is shown in Equation 7.

$$F1score = 2\frac{Precision \times Recall}{+} \quad (7)$$

5.2 Experimental Results and Discussion

After balancing the target classes to 128 sample per each class, and splitting the dataset to three parts, the data enter the model, and exceeded to the training process.

The proposed model which is based on random forest techniques is applied to the processed data. The values of evaluation metrics are as following: the value of accuracy is 100%, 98.11, and 100% for training, testing, and validation, respectively. For Precision: 98.20%, Recall: 98.11%, and finally F1 Score: 98.12%. Figure 2 shows the accuracy model, while the confusion matrix is shown in Figure 3.

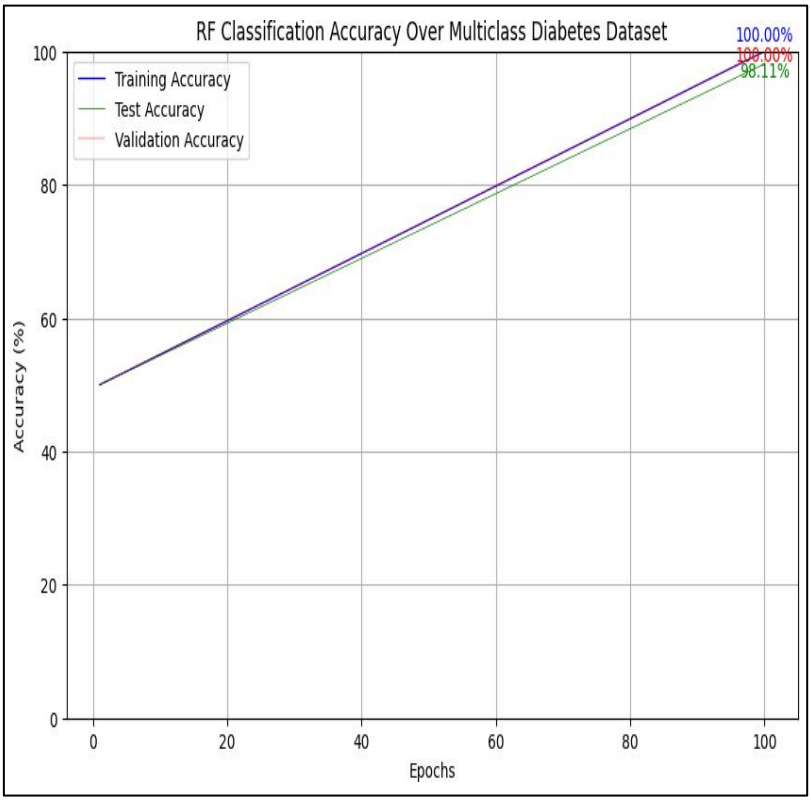


Figure 2: The accuracy of the proposed ensemble model with the multiclass diabetes dataset.

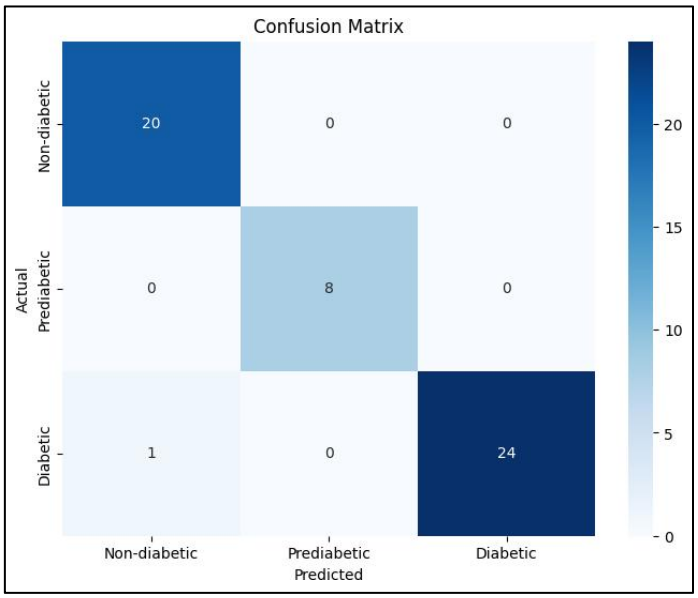


Figure 3: Confusion matrix of the proposed ensemble model with the multiclass diabetes dataset.

While with the multiclass diabetes dataset, the other machine learning methods give values of evaluation metrics as shown in Table 3.

Table 3: The performance metrics of using machine learning methods with the multiclass diabetes dataset.

Metrics	SVM	L R	L R with L1	L R with L2	K-NN	Naïve Bayes
Accuracy	**96.30%**	**92.59%**	84.91%	90.57%	**88.89%**	**96.30%**
Precision	90.57%	90.61%	89%	92%	78.74%	90.85%
Recall	90.57%	90.57%	**96%**	**96%**	79.25%	90.57%
F1 score	90.57%	90.55%	92%	94%	78.78%	90.78%

The differences among these values in Table 3 are shown in Figure 4.

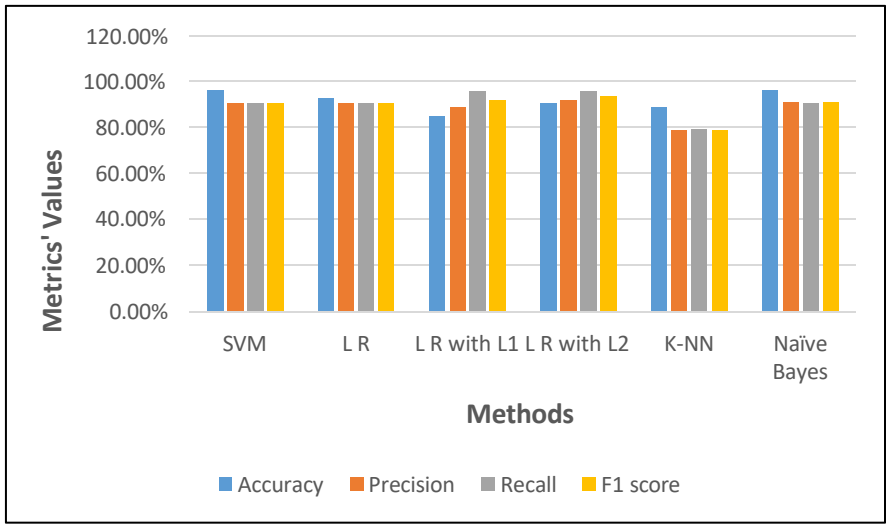


Figure 4: The values of evaluation metrics for each machine learning method that used based on the multiclass diabetes dataset.

While the results of the proposed model when applied to the PIMA dataset are given an accuracy of 85.23%, while the values of other metrics are 88.07% for precision, 87.50% for each of recall and f1 score. Figures 5 and 6 display the accuracy of the proposed system and the confusion matrix, respectively, using the PIMA dataset, where 0 and 1 represent non diabetes and diabetes, respectively.

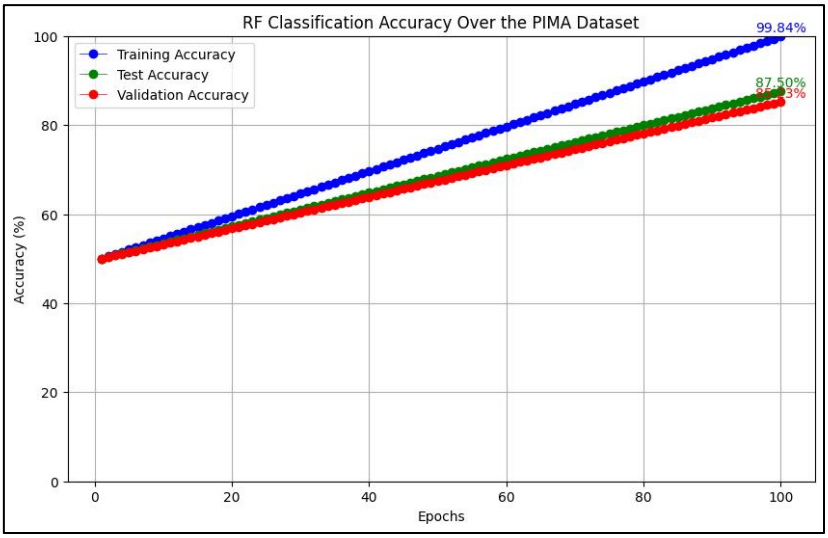


Figure 5: The accuracy of the proposed ensemble model with PIMA diabetes dataset.

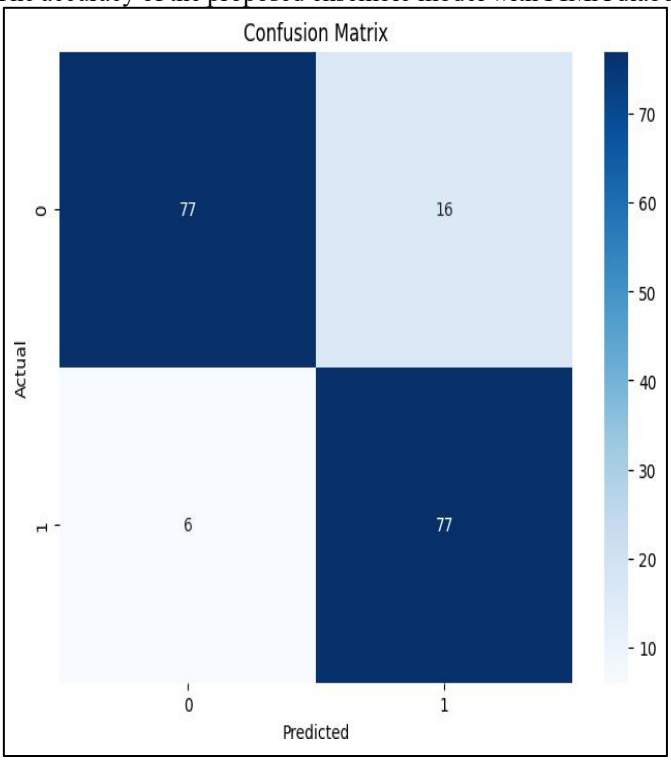
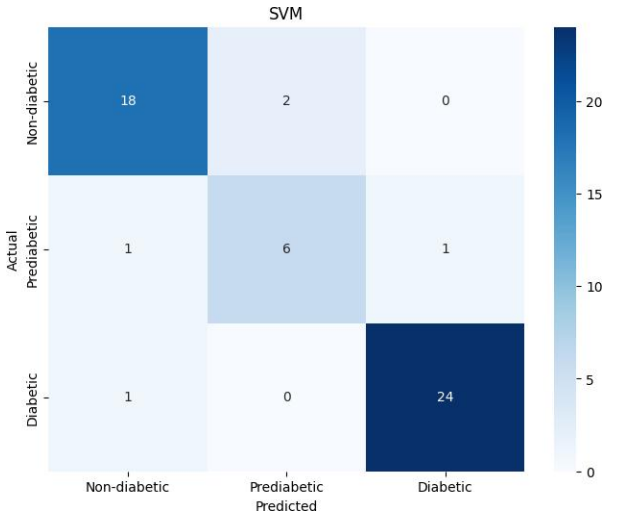


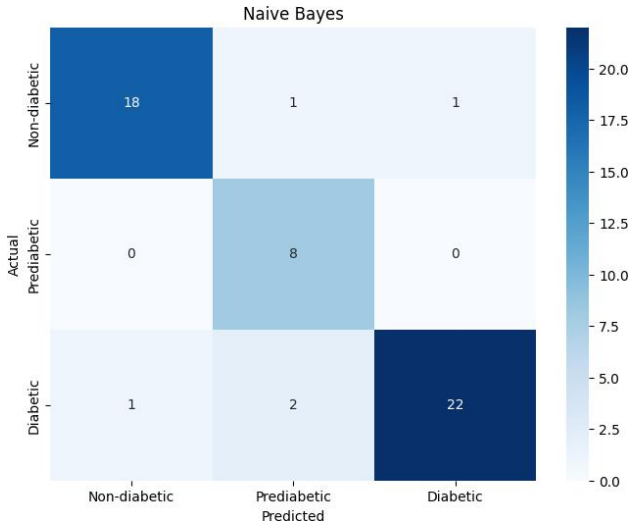
Figure 6: The confusion matrix of the proposed system using PIMA diabetes dataset.

The performance of the proposed model which is based on random forest algorithm return the best classification results since the random forest is an ensemble learning method that used the voting technique and gives higher weights to the more important features which in turn lead to the best performance.

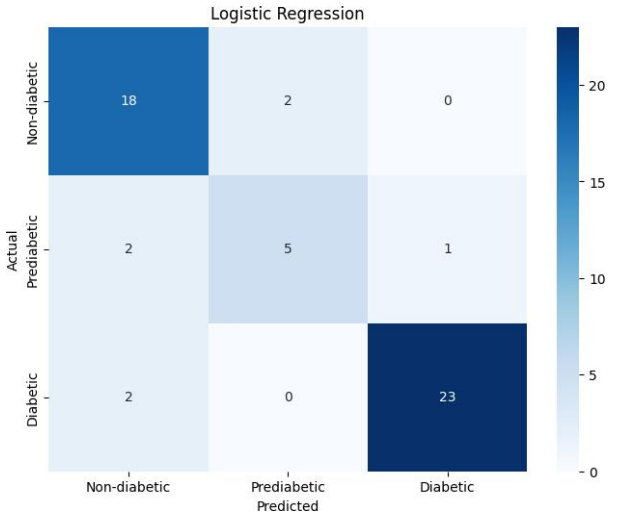
While the other methods achieved less than the proposed model, the confusion matrix with all these machine learning method over the multiclass diabetes dataset is shown in Figure 7. The SVM focus detecting the hyperplane among the target classes and since the classes are good separated, therefore, SVM has a high accuracy value, while the great performance of logistic regression appears with binary classes, and this is the reason why logistic regression achieves good performance but not the best. In logistic regression, feature engineering is necessity, thus, dataset with strong feature engineering will not give promising results as expected, and by applying the PCA plot of multiclass diabetes dataset as shown in Figure 8, it is clear that the boundaries between classes are not entirely linear.



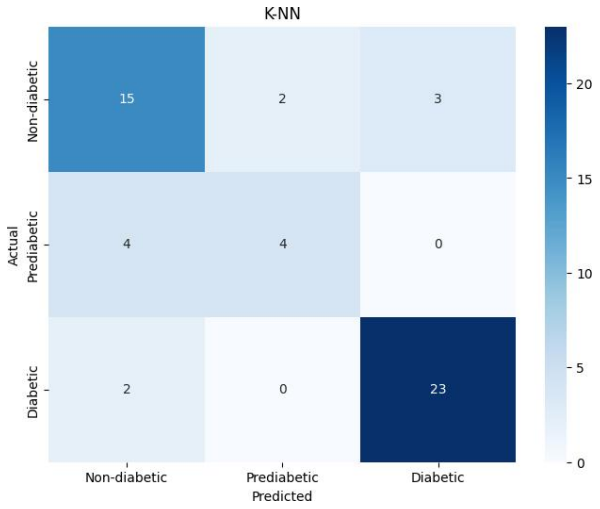
Confusion matrix with SVM



Confusion matrix with naïve bayes



Confusion matrix with Logistic Regression



Confusion matrix with K-NN

Figure 7: The confusion matrices for the used machine learning techniques with multiclass diabetes dataset.

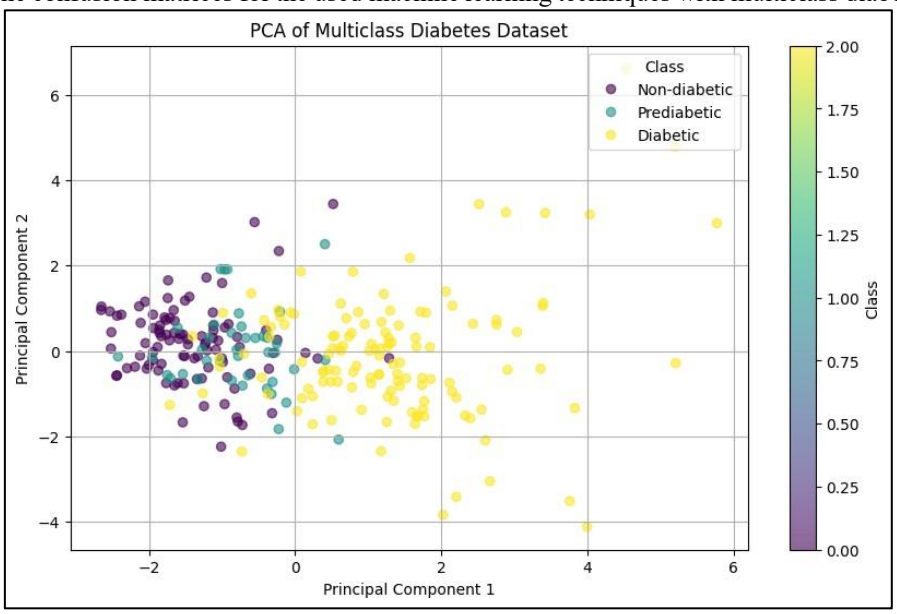


Figure 8: The PCA plot of the multiclass diabetes dataset.

Where there are three colors, each color represents a single class of the three target classes. Thus, it is obvious that the only the yellow color is not completely separated while the two other colors have wide overlapped and this indicates that achieving high accuracy by the models that required separated classes is a challenge and very difficult, and this interpret why logistic regression and logistic regression withL1/L2 regularization do not achieve high accuracy like random forest and even SVM. The naïve bayes is a probabilistic method in which the correlation between features is not preferred and is not lead to the highest performance. Finally, the K-NN classifier does not provide high performance results since the difficulty to choose the optimal value of k, where the smallest or largest value of k cause misclassification of the data. In the case of Pima diabetes dataset, Figure 9 shows the PCA plot of the class on this dataset.

It is obvious that there is an interference between the two classes of the dataset, therefore only ensemble learning techniques can deal with this non-linearity, where the RF is achieved 85.23% accuracy.

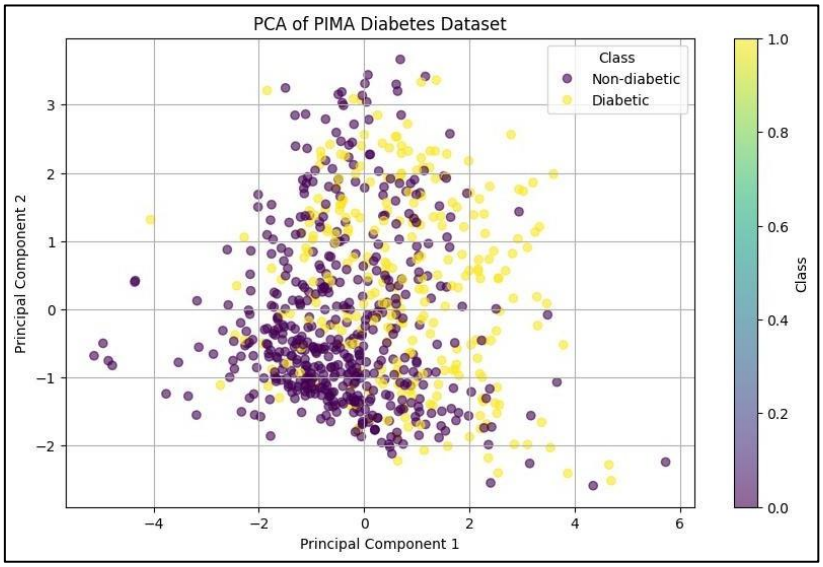


Figure 9: The PCA plot of the Pima diabetes dataset.

6. Conclusion

Diabetes is a major health concern worldwide, with early prediction being crucial for effective management and improving patient outcomes. This study proposed a novel model utilizing the ensemble learning technique to predict diabetes using two datasets, which are Multiclass Diabetes dataset and PIMA Indian Diabetes dataset. With Multiclass Diabetes dataset, the model of RF is achieved outstanding results, including a validation accuracy of 100%, precision of 98.20%, and recall and F1 scores of 98.11% and 98.12%, respectively. While with PIMA Indian Diabetes dataset, the model also demonstrated high performance, with a validation accuracy of 85.23%, precision of 88.07%, and recall and F1 scores of 87.50%. The comparative analysis showed that the random forest algorithm outperformed other machine learning techniques such as SVM, logistic regression, and K-NN. The ensemble learning approach of random forest, which aggregates the decisions of multiple trees, proved effective in handling the complexities of the dataset and class imbalances. The results affirm that ensemble learning is a robust and reliable method for predicting diabetes, capable of providing early diagnosis with high accuracy. This research highlights the potential of machine learning algorithms in healthcare, particularly for diseases with high prevalence and severe outcomes like diabetes. Future work could focus on integrating additional features and datasets to further enhance prediction accuracy and exploring the application of other advanced machine learning techniques. Although the proposed model performs well, it still has some limitations. For instance, the multiclass diabetes dataset has a limited number of instances. Increasing the number of instances would strengthen the dataset, making it more generalizable and able to work efficiently with other models. In contrast, the Pima dataset is a binary classification dataset, which limits its ability to handle multiclass data.

The future direction involves combining multiple algorithms and applying voting or stacking ensemble methods to evaluate the impact of each algorithm and how they can enhance the model's performance.

References

[1] “World Health Organization.” Accessed: Jun. 01, 2024. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

[2] A. K. Nawar et al., “Heart Attack Prediction by Integrating Independent Component Analysis with Machine Learning Classifiers,” Res. Sq., 2024, doi: https://doi.org/10.21203/rs.3.rs-5256555/v1.

[3] L. R. Al-Khazraji, A. R. Abbas, and A. S. Jamil, “A Systematic Review of Deep Dream,” Iraqi J. Comput. Commun.

Control Syst. Eng., vol. 23, no. 2, pp. 192–209, 2023, doi: 10.33103/uot.ijccce.23.2.15.
[4] A. Z. Mohammed and L. E. George, "Osteoporosis detection using convolutional neural network based on dual-energy X-ray absorptiometry images," Indones. J. Electr. Eng. Comput. Sci., vol. 29, no. 1, pp. 315–321, 2022, doi: 10.11591/ijeecs.v29.i1.pp315-321.
[5] L. R. Ali, S. A. Jebur, M. M. Jahefer, and B. N. Shaker, "Employing Transfer Learning for Diagnosing COVID-19 Disease," Int. J. online Biomed. Eng., vol. 18, no. 15, pp. 31–42, 2022, doi: https://doi.org/10.3991/ijoe.v18i15.35761.
[6] L. R. Al-Khazraji, A. R. Abbas, and A. S. Jamil, "The Effect of Changing Targeted Layers of The Deep Dream Technique Using VGG-16 Model," Int. J. online Biomed. Eng., vol. 19, no. 3, pp. 34–47, 2022.
[7] S. Sumathi, S. Rajappa, L. A. Kumar, and S. Paneerselvam, Advanced decision sciences based on deep learning and ensemble learning algorithms: a practical approach using python. Nova Science Publishers, Inc., 2021. doi: https://doi.org/10.52305/XSMY1504.
[8] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and polSAR data: A comparative evaluation," Remote Sens., vol. 13, no. 21, 2021, doi: 10.3390/rs13214405.
[9] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," Int. J. Pharm. Res., vol. 12, no. 4, pp. 56–66, 2020, doi: 10.31838/ijpr/2020.12.04.013.
[10] S. Abdulazeez, A. K. Nawar, N. B. Hassan, and E. Tariq, "Internet of Things: Architecture, Technologies, Applications, and Challenges," AlKadhim J. Comput. Sci., vol. 2, no. 1, pp. 36–52, 2024.
[11] B. S. Zynal, A. T. Lateef, S. A. Jebur, and H. Naser, "Improving Communication Performance Through Fiber Amplifier EDFA," AlKadhim J. Comput. Sci., vol. 2, no. 2, pp. 1–9, 2024.
[12] Z. F. Hussain et al., "A new model for iris data set classification based on linear support vector machine parameter's optimization," Int. J. Electr. Comput. Eng., vol. 10, no. 1, pp. 1079–1084, 2020, doi: 10.11591/ijece.v10i1.pp1079-1084.
[13] M. J. Hazar, B. N. Shaker, L. R. Ali, and E. R. Alzaidi, "Using received strength signal indication for indoor mobile localization based on machine learning technique," Webology, vol. 17, no. 1. pp. 30–42, 2020. doi: 10.14704/WEB/V17I1/A206.
[14] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," BMC Bioinformatics, vol. 24, no. 1, pp. 1–20, 2023, doi: 10.1186/s12859-023-05467-x.
[15] "Multiclass Diabetes Dataset." Accessed: May 30, 2024. [Online]. Available: https://data.mendeley.com/datasets/jpp8bsjgrm/1
[16] "National Institute of Diabetes and Digestive and Kidney Diseases (2022) Pima Indians Diabetes - dataset by uci | data.world." Accessed: May 30, 2024. [Online]. Available: https://data.world/uci/pima-indians-diabetes
[17] L. R. Ali, H. K. Homood, and A. S. Elameer, "Feature Extraction Techniques on Facial Images : An Overview," Int. J. Sci. Res., vol. 6, no. 9, pp. 2015–2018, 2017, doi: 10.21275/ART20176682.
[18] M. Phongying and S. Hiriote, "Diabetes Classification Using Machine Learning Techniques," Computation, vol. 11, no. 5, 2023, doi: 10.3390/computation11050096.
[19] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," Diagnostics, vol. 13, no. 14, 2023, doi: 10.3390/diagnostics13142383.
[20] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," IEEE World Forum Internet Things, WF-IoT 2018 - Proc., vol. 2018-Janua, no. July, pp. 291–295, 2018, doi: 10.1109/WF-IoT.2018.8355130.
[21] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/9930985.
[22] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and prediction of diabetes disease using decision tree method," CEUR Workshop Proc., vol. 2824, pp. 163–172, 2021.
[23] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in Ensemble machine learning: Methods and applications, Springer Science & Business Media, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7.
[24] P. T. Nguyen et al., "Soft computing ensemble models based on logistic regression for groundwater potential mapping," Appl. Sci., vol. 10, no. 7, p. 2469, 2020, doi: 10.3390/app10072469.
[25] M. Awad and R. Khanna, "Support Vector Machines for Classification," in Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 2015, pp. 39–66. doi: 10.1007/978-1-4302-5990-9.
[26] D. Berrar, "Bayes' theorem and naive bayes classifier," in Encyclopedia of Bioinformatics and Computational Biology, vol. 1, 2019, pp. 403–412. doi: 10.1016/B978-0-12-809633-8.20473-1.
[27] S. A. Salman, A.-H. Ayad Salih, A. Hussein Ali, M. Khamees Khaleel, and M. Abdulghfoor Mohammed, "A New Model for Iris Classification Based on Naïve Bayes Grid Parameters Optimization," Artic. Int. J. Sci. Basic Appl. Res., vol. 40, no. 2, pp. 150–155, 2018, [Online]. Available: http://gssrr.org/index.php?journal=JournalOfBasicAndApplied
[28] S. Zhang, "Challenges in KNN Classification," IEEE Trans. Knowl. Data Eng., vol. 34, no. 10, pp. 4663–4675, 2022,

doi: 10.1109/TKDE.2021.3049250.
[29] M. Awad and R. Khanna, Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Springer Nature, 2015.
[30] J. Qin and Y. Lou, "L1-2 Regularized Logistic Regression," in 2019 53rd Asilomar conference on signals, systems, and computers, IEEE, 2019, pp. 779–783. doi: 10.1109/IEEECONF44664.2019.9048830.
[31] S. A. Jebur, K. A. Hussein, and H. K. Hoomod, "Abnormal Behavior Detection in Video Surveillance Using Inception-v3 Transfer Learning Approaches," IRAQI J. Comput. Commun. Control Syst. Eng., vol. 23, no. 2, pp. 210–221, 2023.