A Workflow Paper on Prediction of Type 2 Diabetes Using Ensemble Learning Method

[1]Arwinder Singh, [2]Neeraj Sharma, [3]Nishita Gouraha, [4]Ruchita Yadav, [5]Aparna Pandey

[1,2,3,4]Student, [5]Assistant Professor
Department of Computer Science and Engineering
Bhilai Institute of Technology
Raipur, Chhattisgarh, India

Abstract - **Diabetes is a type of chronic disease that develops from lack of insulin in our body. In diabetes, this process is broken. The main forms of diabetes are type 1 and type 2, but there are other forms as well, including gestational diabetes, which develops during pregnancy. The use of various Machine Learning algorithms including K-Nearest Neighbors (KNN), random forests (RF), decision trees (DT), AdaBoost (AB), Naive Bayes classifier (NB), and XGBoost (XB), and preprocessing steps includes outlier rejection, filling in missing values, data standardization, and stratified K-fold validation to validate the results. To enhance the outcome, the weighted ensembling of various machine learning models are also suggested. For performance metric Area Under ROC Curve (AUC) is used. For further optimization in model's performance is done using Grid Search technique of hyperparameter tuning. In a publicly accessible Pima Indian Diabetes Dataset from Kaggle in which 768 female patients record is given and 268 are diabetic and 500 are non-diabetic.**

Keywords - **Diabetes Prediction; Ensemble Learning; Random Forest; Accuracy; AUC; K-Nearest Neighbour; Decision Tree; XGBoost; Naïve Bayes Classifier; Outlier Removal; PCA**

1. INTRODUCTION

Diabetes mellitus (DM) is considered as a chronic disease that has been affecting people of all age groups [1]. According to health experts, diabetes occurs when the human body's gland called the pancreas cannot produce enough insulin (Type 1 diabetes), and the produced insulin cannot be used by the cell of the body (Type 2 diabetes) [2]. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves [3]. There is no permanent cure for diabetes [2]. Automated identification with better accuracy is essential for the early detection of diabetes [4].

Diabetes exists in three forms :- **Diabetes Mellitus Type-1** is characterized by pancreas generating insulin less than what is required by the body, a condition also called "insulin-subordinate diabetes mellitus" (IDDM). People suffering from type-1 DM require external insulin dosage to make up for the less insulin produced by the pancreas [5]. **Diabetes Mellitus Type-2** is marked by the body resisting insulin as the body cells react differently to insulin than they normal would. This may ultimately lead to no insulin in the body. This is otherwise called "non-insulin subordinate diabetes mellitus" (NIDDM) or "adult starting diabetes". This type of diabetes is commonly found in people with high BMI or those who lead an inactive lifestyle [5]. **Prediabetes** – A person is considered to have prediabetes if body glucose concentration is 100 to 125 mg/dl [2]. The "gestational diabetes (GDM)" occurs mostly during pregnancy [1].

A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events [6]. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made [6].

About 422 million people worldwide have diabetes, the majority living in low-and each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades [3]. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes [3]. There is no doubt that this alarming figure needs great attention [3]. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health [3]. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws [6]. This project aims to identify the most accurate machine learning method that can determine whether or not a person has diabetes.

The paper is organised into V sections. Section I introduces diabetes prediction, including types of diabetes. Section II discusses literature reviews; and Section III presents a proposed methodology in detail. Section IV presents results and analysis of the proposed methodology. In section V, the conclusion has been given.

2. LITERATURE REVIEW

Priyanka Goyal, Somil Jain in [7] used machine learning classification algorithms - Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest) for the prediction of type 2 diabetes and uses PIMA Indian Diabetes Database (PIDD) dataset, accuracy of ensemble Method is 77.60%.

Jana S, Bharanidharan N, ShanmukhaNagasai P, Saravan Kumar K, Mani Nageshwar V in [8] applied machine learning algorithms (i.e., K-nearest Neighbour, Decision Tree and Random Forest) with the accuracy of 77%. The information was acquired from the Pima Indian disease Dataset (PIDD).

Sourav Kumar Bhoia , Sanjaya Kumar Pandab , Kalyan Kumar Jenaa , P. Anshuman Abhisekhc , Kshira Sagar Sahood , Najm Us Samae,* , Shweta Supriya Pradhan c , Rashmi Ranjan Sahoo in [9] used supervised learning methods such as classification tree

(CT), support vector machine (SVM), kNearest Neighbour (k-NN), Naïve Bayes (NB), Random Forest (RF), Neural Network (NN), AdaBoost (AB)and Logistic Regression (LR). . In this study, the developed model uses the clinical dataset to forecast the diabetes in female Pima Indians. Logistic Regression (LR) was found to be with 76.8% accuracy.
Ram D. Joshi Chandra K. Dhakal in [10] predicted type 2 diabetes for Pima Indian women utilizing a logistic regression model and decision tree—a machine learning algorithm. Preferred specification yields a prediction accuracy of 78.26%.
Neha Prerna Tigga , Shruti Garg in [5] predicted The risk of Type 2 diabetes using different machine learning algorithms-Logistic Regression Method , KNN , SVM , Naïve Bayes Classification Method , Random Forest Classification. . 952 instances have been collected through an online and offline questionnaire and PIMA dataset was also used. The performance of Random Forest Classifier is found to be most accurate for both datasets.
N. Sneha Tarun Gangil in [11] used - Decision tree, Naïve Bayesian, Support vector machine, Random forest, K nearest neighbour (KNN). The PIMA dataset is collected from UCI machine repository. The result shows the accuracy of SVM is 77.73%. The accuracy of random forest is 75.39.
Perveen et. al. in [12] Naive Bayes and the J48 (C4.5) decision tree model were used for diabetes prediction and k-medoids sampling was used to balance the training set. NB did better than the others in their study.
Zou et. al. in [3] Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods were used for feature selection methods. Random Forest, J48, and ANN were used for categorization. It is found that mRMR accuracy is superior to PCA with all attributes.
S. Saru, S. Subashree [13], applied 10-fold cross validation and ensemble method. The highest accuracy was given by Decision Tree -94.4%. Author concluded that ensemble method gives better performance than single method.
Saloni Kumari, Deepika Kumar, et.al. [14], proposed an approach of combining three machine learning algorithms like Random Forest, Logistic Regression, and Naive Bayes. Soft voting classifier was employed for ensembling the models. The PIMA Indian diabetes dataset was utilized and the accuracy of 79.02% was achieved.
Messan Komi, Jun Li, et.al. [15], applied Artificial Neural Network (ANN), Gaussian Mixture Models (GMM), Support Vector Machine (SVM), Logistic regression, ELM, for diabetes prediction and ANN gave the highest accuracy among all.
R. Karthikeyan, Dr. P.Geetha, et.al. [16], has predicted diabetes using rule-based classifier and decision tree J48 with missing values detection on PIMA dataset and accuracy given by model is 95.12%.
Huma Naz, Sachin Ahuja [17], have proposed an approach where they have used a concept of Deep Learning Classifiers and achieved accuracy of 98.07%.
Maham Jahangir, Hummad Afzal, et.al. [18], used concept of Automatic Multilayer Perceptron (AMP) on PIMA diabetes dataset with outlier removal and gave accuracy of 88.7%.

3. MATERIALS AND METHODS

3.1 Dataset Used

Table I. Attributes in Dataset

	Attributes	Description
F1	GLUCOSE	Plasma glucose concentration over 2h in an oral glucose tolerance test.
F2	PREGNANCIES	It shows how many times patient is pregnant.
F3	BLOOD PRESSURE	It indicates the BP of patient.
F4	SKIN THICKNESS	It shows skin fold thickness.
F5	DIABETES PEDIGREE FUNCTION	It shows family history of patient.
F6	BMI	It indicates Body Mass Index.
F7	INSULIN	2-Hour serum insulin (mu U/ml).
F8	AGE	It shows age of patient. The age group to be used is 21-81 for analysis.
F9	OUTCOME	1 for diabetes and 0 for non-diabetes.

The dataset to be utilized is PIMA Indian Diabetes Dataset (PIDD). The data is collected from Kaggle. This dataset is mainly used to predict whether a person has diabetes or not. The dataset contains details of 768 patients and their corresponding nine unique attributes.
The nine attributes that are used for this paper are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome. The 'outcome' attribute is taken as a dependent variable and the remaining eight attributes are taken as independent variables. The diabetes attribute 'outcome' consists of binary value where 0 means non-diabetes, and 1 means diabetes. The nine attributes in the dataset are described in Table I [19]
The Pedigree (Diabetes Pedigree Function) was calculated in [20]:

$$\text{Pedigree} = \frac{\sum \text{iKi}(88 - \text{ADMi}) + 20}{\sum \text{jKj}(\text{ALCj} - 14) + 50}$$

where i and j respectively denote the relatives who had developed and NOT developed diabetes. K is the percentage of shared genes by the relatives (K=0.500 for the parent or full sibling, K=0.250 for a half-sibling, grandparent, aunt or uncle and K=0.125 for a half-aunt, half-uncle, or first cousin). ADMi and ACLj are the age of relatives, in years, at the time of diagnosis and the last non-diabetic test [20].

3.2 Proposed Framework:

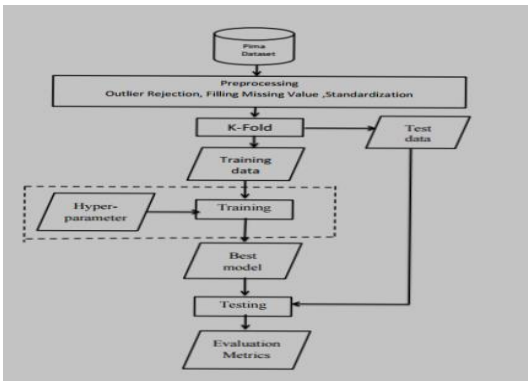

Figure 1. Flowchart of the proposed methodology

The proposed framework, in this literature, has been illustrated in the fig 1. We have used Python Programming Language for coding and platform used is Jupyter Notebook. The first step is to collect the dataset which we have collected from Kaggle and the name of the dataset is PIMA Indian Diabetes Dataset (PIDD). Then three types of data pre-processing are done. Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. Data Pre-processing helps us to reduce the noise and fill out missing values to maximize the performance of a machine learning algorithm. Three steps in this data pre-processing are Outlier Rejection, Missing values removal and Standardization. Outliers are values extremely distinct from other data points Therefore to remove them outlier rejection is done. Standardization is the method to bring all data points at one level or scale, also known as Z-score. Standardization is another scaling method where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation [21]. Three types of feature selection have been applied -Principal Component Analysis (PCA), Independent Component Analysis. Correlation based feature selection. The classification algorithms utilized are – Logistic Regression, AdaBoost, XGBoost Classifier, Decision Tree, Random Forest, K-Nearest Neighbour. To validate the data stratified k-cross fold validation method is implemented. Ensemble learning is implemented in which several models are combined to get better results. Voting Classifier is used to perform ensemble learning and is described as A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator The aggregating criteria can be combined decision of voting for each estimator output. The voting criteria can be of two types: Hard Voting: Voting is calculated on the predicted output class. Soft Voting: Voting is calculated on the predicted probability of the output class [22]. Total five combinations are implemented described in results section of the paper. To maximize the performance of the model Hyperparameter tuning is applied and performance of all combinations of models is evaluated and compared using evaluation metrics explained further in the paper.

3.3 Brief description of Machine Learning Classification Techniques

3.3.1. Logistic Regression Method:

Logistic Regression [23] is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

3.3.2. AdaBoost Method:

AdaBoost [24] (**Ada**ptive **Boost**ing) is a very popular boosting technique that aims at combining multiple weak classifiers to build one strong classifier. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees re also called **Decision Stumps.**

XGBoost [25] is an optimized distributed gradient boosting library designed to be highly efficient**,** flexible and portable**.**It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

3.3.3. Decision Tree Method

A decision tree [26] is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. It is a very specific type of probability tree that enables you to make a decision about some kind of process.

3.3.4.KNN Method:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. K-Nearest Neighbors (KNN) is a standard machine-learning method that has been extended to large-scale data mining efforts. The idea is that one uses a large amount of training data, where each data point is characterized by a set of variables [27].

3.3.5. Random Forest Classifier:

Random forests [28] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large.

3.3.6 Naïve Bayes Classifier:

Naïve bayes [29] classification method is a probabilistic machine learning algorithm based on Bayes theorem described in probability. Even with its simplicity it outperforms other classifiers: hence, it is one of the best classifiers. The Bayes theorem for calculating posterior probability is given below

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)P(c)}{P(x)}$$

$$P\left(\frac{c}{x}\right) = \mathrm{P}\left(\frac{\mathrm{x1}}{\mathrm{c}}\right) \times \mathrm{P}\left(\frac{\mathrm{x2}}{c}\mathrm{c}\right) \times \ldots \times \mathrm{P}\left(\frac{\mathrm{xn}}{c}\right) \times \mathrm{P(c)}$$

Where,

P(c / x) = posterior probability of class (target) given predictor • (attribute).

P(c) = prior probability of class.

P(x / c) = probability of predictor given class.

P(x) = probability of predictor.

3.4 Evaluation Metrics:

Evaluation metrics are used to measure the quality of the statistical or machine learning model Evaluating machine learning models or algorithms is essential for any project There are many different types of evaluation metrics available to test a model These include classification accuracy, logarithmic loss, confusion matrix, and others [30].

Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data [31]. Figure 2. taken from [32].

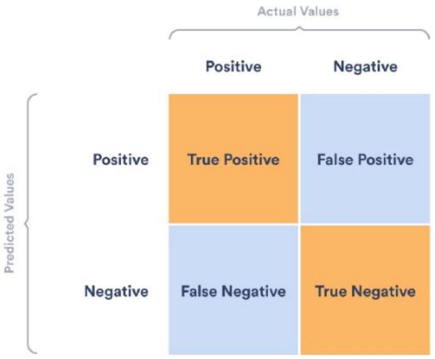

Figure 2. Confusion Matrix Example

True Positive (TP): It refers to the number of predictions where the classifier correctly predicts the positive class as positive [33].
True Negative (TN): It refers to the number of predictions where the classifier correctly predicts the negative class as negative [33].

False Positive (FP): It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive [33].
False Negative (FN): It refers to the number of predictions where the classifier incorrectly predicts the positive class as negative [33].

Accuracy, Precision, Recall, Specificity is described in Table II [33]. AUC-ROC is also explained in Table II [34].

Table II: Description of Evaluation Parameters

S. No	Measures	Definition	Formula
1.	Accuracy	Accuracy is the fraction of predictions our model got right.	$A = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$
2.	Precision	It tells you what fraction of predictions as a positive class were actually positive.	$P = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$
3.	Recall	It tells you what fraction of all positive samples were correctly predicted as positive by the classifier.	$R = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$
4.	Specificity	It tells you what fraction of all negative samples are correctly predicted as negative by the classifier.	$Specificity = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})}$
5.	AUC-ROC	ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.	

4. RESULT

Table III. Model Combination

S.NO	Ensemble Model
1	AB+XB
2	k-NN+DT+XB
3	DT+AB+RF+XB
4	k-NN+DT+RF+XB+NB
5	k-NN+DT+RF+XB+NB+AB

Table IV: Evaluation of model combination

Evaluation Parameters	1st Combination	2nd Combination	3rd Combination	4th Combination	5th Combination
AUC	0.946	0.943	0.945	0.940	0.940
Accuracy	0.903	0.888	0.889	0.895	0.893
Sensitivity	0.793	0.808	0.818	0.813	0. 813
Specificity	0.952	0.925	0.936	0.932	0. 929
Precision	0.882	0.832	0.855	0.845	0. 841

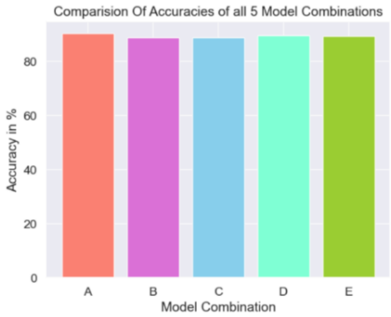

Figure 3. Visual Representation of Comparison of Accuracies of all 5 Model Combination

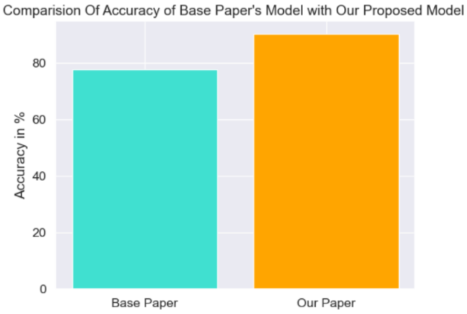

Figure 4. Visual Representation of Comparison of Accuracies of Base Paper and Our Paper

We have selected [7] as our base paper and it gives accuracy of 77.60% that is way more lesser than the accuracy our best model is giving that is 90.3%.

5. CONCLUSION

In this paper, a new approach is proposed by combining various classification algorithms. We have used Logistic Regression, AdaBoost, XGBoost Classifier, Decision Tree, Random Forest, K-Nearest Neighbour for this research paper. The accuracy for our dataset (PIMA dataset) given by our best model combination (AdaBoost and XGBoost) is 90.3% which is better than other papers. AUC is maximized to 0.946 which is considered to be great. Hence, we can say that this methodology is really useful for early detection of diabetes. This approach can also be used to predict other diseases. For future, this approach can be enhanced by using any other technique of Machine Learning as machine learning is a vast field. Any other algorithm can be applied to improve the work.

REFERENCES:

[1] M. S. A. A. O. N. V. B. Shamreen Ahamed, "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," *Frontiers Computer Science,* vol. 4, 2022.

[2] S. Y. F. Jobeda Jamal Khanam, "A comparison of machine learning algorithms for diabetes prediction," vol. 7, no. 4, pp. 432-439, 2021.

[3] K. Q. Y. L. D. Y. Y. J. H. T. Quan Zou, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Gene,* vol. 9, 2018.

[4] M.-C. C. ,. C.-J. W. Cheng-Lung Huang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications,* vol. 33, no. 4, pp. 847-856, 2007.

[5] S. G. Neha Prerna Tigga, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Computer Science,* vol. 167, pp. 706-716, 2020.

[6] N. Y. K. Y. &. P. U. P. Sarvesh Wadekar, "Diabetes Prediction System Using Gaussian Algorithm.," *International Journal of Advanced Research in Science, Communication and Technology,* pp. 29-34, 2022.

[7] P. Goyal and S. J. , "Prediction of type-2 diabetes using classification and ensemble method approach," in *International Mobile and Embedded Technology Conference (MECON),* Noida, India, 2022.

[8] J. S, B. N, S. P, S. K. K and V. M. Nageshwar, "Diabetes Prediction Using Machine Learning Algorithms," in *8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, , 2022.

[9] e. a. Sourav Kumar Bhoi, "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach," *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* vol. 12, pp. 3074-3084, 2021.

[10] C. K. D. Ram D. Joshi, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *International journal of environmental research and public health,* 2021.

[11] T. G. N. Sneha, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data,* 2019.

[12] M. S. K. K. a. A. G. S. Perveen, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques," *IEEE Access,* vol. 7, pp. 1365-1375, 2019.

[13] S. S. S. Saru, "Analysis and Prediction of Diabetes Using Machine Learning," *International Journal of Emerging Technology and Innovative Engineering,* vol. 5, no. 4, p. 9, 2019.

[14] D. K. M. M. Saloni Kumari, "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.

[15] M. Komi, J. Li, Y. Zhai and X. Zhang, "Application of data mining methods in diabetes prediction," in *2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, 2017.

[16] R. Karthikeyan, P. Geetha and E. Ramaraj, "Rule Based System for Better Prediction of Diabetes," in *3rd International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, India, 2019.

[17] S. A. Huma Naz, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders* , p. 391–403, 2020.

[18] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," in *Intelligent Systems Conference (IntelliSys)*, London, UK, 2017.

[19] M. B. Rashi Rastogi, "Diabetes prediction model using data mining techniques," *Measurement: Sensors,* vol. 25, 2023.

[20] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access,* vol. 8, pp. 76516-76531, 2020.

[21] A. Bhandari, Analytics Vidhya, 3 April 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.

[22] Towards Data Science, 28 October 2021. [Online]. Available: https://towardsdatascience.com/use-voting-classifier-to-improve-the-performance-of-your-ml-model-805345f9de0e.

[23] [Online]. Available: https://www.statisticssolutions.com/.

[24] R. E. Schapire, "Explaining AdaBoost," *Empirical Inference* , p. 37–52, 2013.

[25] [Online]. Available: https://xgboost.readthedocs.io/en/stable/.

[26] [Online]. Available: https://www.statisticshowto.com/decision-tree-definition-and-examples/.

[27] P. Nadkarni, Clinical Research Computing: A Practitioner’s Handbook, Academic Pfress.

[28] L. Breiman, "Random Forests," *Machine Learning* , p. 5–32, 2001.

[29] [Online]. Available: https://www.saedsayad.com/naive_bayesian.htm.

[30] [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics.

[31] [Online]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/.

[32] [Online]. Available: https://plat.ai/blog/confusion-matrix-in-machine-learning/.

[33] [Online]. Available: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826.

[34] [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.