The Effect of Combining Datasets in Diabetes Prediction Using Ensemble Learning Techniques

Emad Majeed Hameed[1*], Hardik Joshi[2], and Ahmed Abdul Azeez Ismael[3]
[1,2]Department of Computer Science, Gujarat University
Gujarat, India 380009
[1,3]Baquba Technical Institute, Middle Technical University
Baghdad, Iraq 10074
Email: [1]emadhameed@gujaratuniversity.ac.in, [1]emadmajeed@mtu.edu.iq,
[2]hardikjoshi@gujaratuniversity.ac.in, [3]Ahmed.abdulazeez@mtu.edu.iq

***Abstract*—Diabetes prediction models often suffer from limited generalizability due to reliance on single-population datasets, which fail to capture the diversity of real-world patient demographics. This limitation reduces their clinical applicability across different ethnic groups and geographic regions. The research aims to improve diabetes prediction accuracy and generalizability by combining multiple datasets and employing ensemble learning techniques, addressing the challenges of imbalanced data and population diversity. The research combines two publicly available datasets (Pima Indians: 768 samples and German Society: 2,000 samples) and utilizes preprocessing procedures conducted on these datasets. By comparing the performance of the individual dataset (Pima Indians and German Society datasets) and the combined dataset, it is clear that the models trained on the combined data show improved performance on all metrics. The Random Forest model outperforms the other ensemble models in the Pima Indians dataset, achieving an accuracy of 0.817. The models with the highest accuracy on the German Society dataset are Gradient Boosting and Random Forest, with respective accuracies of 0.996 and 0.994. Then, in the combined dataset, Gradient Boosting and Random Forest yield the best accuracy of 0.991 and 0.988, respectively. It is noticeable that this improvement reflects the ability of models trained on combined data to better accommodate diversity in the data, allowing them to generalize more effectively when applied to different populations. Future research should explore deep learning techniques and additional diverse datasets to enhance model performance further.**

***Index Terms*—Combined Dataset, Diabetes Prediction, Ensemble Learning**

Received: Aug. 22, 2024; received in revised form: Dec. 05, 2024; accepted: Dec. 05, 2024; available online: May 08, 2025.
*Corresponding Author

I. Introduction

DIABETES is a chronic disease that occurs when the body does not produce enough insulin. Diabetes can lead to serious complications if left untreated. Uncontrolled diabetes has a negative impact, especially on blood vessels. The main organs and tissues that are permanently damaged by sugar include the heart, brain, leg, kidney, and nerve endings [1, 2]. Early diagnosis of diabetes is of utmost importance and vital to prevent much of the damage. Medical studies have shown that diabetes has worsened recently, with 537 million people currently living with diabetes worldwide, and 6.7 million people died from the disease in 2021 alone [3].

Traditional methods of detecting diabetes by human health experts are done through manual examinations or by examining blood samples taken from patients with the help of a medical device in a laboratory setting. However, since diabetes develops without showing many symptoms, it may not be diagnosed even by doctors who are experts in their field [4]. Technological developments make it possible to diagnose many diseases using smart and learning methods. In this way, the diagnosis of diseases and reporting of relevant examinations are completed in a shorter time. As a result, the time spent by patients in the healthcare institution is reduced [5].

At present, large investments are being made in smart hospital projects in many countries. This application reduces the density of healthcare institutions and the amount of labor required by automating the system. Machine learning and data mining-based methods are of great importance for the detection, management, and other related clinical treatment of diabetes. Early

diagnosis of diabetes is greatly assisted by computer-aided expert systems based on machine learning [6]. Intelligent technologies such as machine learning and data mining offer promising solutions to improve the efficiency of diabetes risk prediction models. These technologies analyze large and complex datasets to uncover hidden patterns and relationships that traditional statistical methods may miss [7]. Intelligent models can provide more accurate assessments for diabetes prediction by leveraging data sources such as demographics, clinical biomarkers, genetics, lifestyle behaviors, and environmental factors. Despite the promising approaches offered by intelligent modeling, several challenges must be addressed to maximize its effectiveness in predicting diabetes risk. These challenges include data quality and availability, model interpretability, and generalizability across diverse populations [8].

Since existing approaches frequently rely on single data sources, which restricts their generalizability, the combination of different datasets for diabetes prediction represents a major research need. Researchers can improve the diversity and accuracy of prediction models by combining multiple datasets, therefore mitigating the biases present in single-source data. Diabetes predictors across different demographics may be better understood, which eventually improves model performance, using this method [9, 10].

The research proposes and evaluates intelligent techniques for diabetes risk prediction to address those challenges. The research aims to improve predictive model accuracy, interpretability, and generalizability by utilizing current machine learning algorithms. Compared to the previous studies in the literature, the researchers combine datasets and various methods used, like the exploratory data analysis stage, to resolve the problem of missing values and determine and fix the outliers values. Imbalances in the dataset are 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 the machine-learning model [11].

A. Literature Review

The health sector is currently one of the most significant applications of technological innovations. Artificial intelligence technologies are among the methods of choice for improving health-related efficiency, timely treatment planning, and accurate and rapid disease diagnosis [6, 12]. The employment 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 are conducted on the dataset known as "Pima Indians diabetes". These studies aim to predict diabetes through data mining techniques. The first research [13] has used 35,669 patient records and 30 attributes from July 2004 to April 2014, taken from the Endocrine Department of a Chinese university, to analyze real-life data for diabetes diagnosis. It is apparent that Adaboost.M1 and LogitBoost yield superior results in terms of computation time and accuracy when applied to the obtained data alongside the Logistic Regression, Random Forest, and Adaboost.M1 approaches. The accuracy of the LogitBoost method is 93.93% higher than that of the Adaboost.M1 approach. Additionally, the proper classification rate of the LogitBoost algorithm reaches 95.30% when the dataset for the Adaboost.M1 and LogitBoost algorithms is divided into training and test data using the 10-fold cross-validation technique. Next, using the Bayesian method, Naive Bayes, J28, Random forest, Random tree, REP, KNN, CART, and associative rule learning algorithms, in the second research, the previous researchers [14] compare the performance of these algorithms on this dataset.

The third research [15] has used the Deep Neural Network methodology, an unsupervised learning method, to diagnose diabetes on the Pima Indians diabetes dataset effectively. It additionally employs a feature selection model with Random Forest and Extra Trees to choose significant features. With an accuracy of 98.16%, the model outperforms other contemporary methods in the field.

Similarly, using the Pima Indians diabetes dataset, in the fourth research, the researchers [16] have investigated a number of machine learning techniques for diabetes diagnosis. Out of all the machine learning techniques used, linear discriminant analysis demonstrates the highest accuracy, at 77%. Then, the fifth research [17] conducts a study that makes use of the same dataset. The diabetes classification is based on machine learning algorithms employed: J.48, Naive Bayes, and Random Forest algorithms. The effect of feature selection on classification models is investigated by looking at the outcomes of models with and without feature selection, relying on three- and five-factor feature selection. Among the algorithms employed in the investigation, the Random Forest algorithm outperforms the other two algorithms and models that undergo feature selection, with an accuracy rate of 79.57% in the absence of feature selection.

The sixth research [18] compares the effectiveness of the Decision Tree, XGBoost, Random Forest, Logistic Regression, AdaBoost, and Support Vector Machine (SVM) in diagnosing diabetes. Among the other models, AdaBoost has the highest accuracy (83%). Then,

TABLE I
KEY STATISTICS OF FEATURES IN THE PIMA INDIANS DATASET, HIGHLIGHTING GLUCOSE AND BODY MASS INDEX (BMI) TRENDS, INDICATING A POPULATION WITH ELEVATED GLUCOSE AND BMI VALUES, BOTH CRITICAL FACTORS IN DIABETES RISK.

Index	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Count	768	768	768	768	768	768	768	768	768
Mean	3.845	120.890	69.105	20.536	79.799	31.992	0.471	33.240	0.348
Std.	3.369	31.972	19.355	15.952	115.240	7.884	0.331	11.760	0.476
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.078	21.000	0.000
25%	1.000	99.000	62.000	0.000	0.000	27.300	0.243	24.000	0.000
50%	3.000	117.000	72.000	23.000	30.500	32.000	0.372	29.000	0.000
75%	6.000	140.250	80.000	32.000	127.250	36.600	0.626	41.000	1.000
Max	17.000	199.000	122.000	99.000	846.000	67.100	2.420	81.000	1.000
Data type	int64	int64	int64	int64	int64	float64	float64	float64	float64

the seventh research [19] focuses on the feature selection that is involved in the early prediction of diabetes. It aims to identify the important features and find the most appropriate machine learning classifier that provides the closest result to clinical results. Decision Tree and Random Forest have the best specificity with 98.20% and 98.00%, respectively. Meanwhile, Naive Bayes offers the best accuracy with 82.30%.

A Deep Neural Network classifier has been developed by the eighth research [20] for classification purposes. With an accuracy rate of 98.16%, the model used outperforms other previous studies. However, they also note that the computation time is the main limitation of the study. Hence, future investigations should focus on optimizing computation time and enhancing its efficiency. SVM and Artificial Neural Network (ANN) are utilized by the ninth research [21] to predict diabetes diagnosis. The model's accuracy is 94.87.

The tenth research also [22] utilizes the Pima Indians diabetes dataset to investigate diabetes classification algorithms. The Decision Tree C4.5 and K-Means clustering techniques are combined to create a hybrid model. The hybrid model, which is run in two stages, provides a higher correct classification rate than the classification rate achieved with only the Decision Tree C4.5 methods.

In order to diagnose diabetes, the eleventh research [23] examines data from the Sylhet Diabetes Hospital in Bangladesh, made accessible through the UCI (the Machine Learning Repository, a well-known online dataset archive for algorithmic empirical research run by California University). There are 520 samples in this dataset, and each sample has 17 features. They employ the 10-fold cross-validation technique during the data splitting phase and the 0-1 normalization technique during the data preprocessing phase. The performance of six distinct machine learning techniques—ANN, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, SVM, and Adaboost—is compared. With an accuracy of 98.1%, 98.4%, and 98.4%, respectively, the suggested neural network model achieves the maximum accuracy, F1-score, and specificity values.

II. RESEARCH METHOD

A. Dataset

The research uses three datasets to predict diabetes: the Pima Indians Diabetes dataset [24], the Frankfurt Hospital, the German Society dataset [25], and a combined dataset of aforementioned datasets. They are available to the general public at the Kaggle machine learning repository. The German Society dataset is very similar to the Pima Indians dataset in terms of features, with the exception that it has a greater number of data points. The Pima dataset has 768 records and 9 features. Meanwhile, the German Society dataset has 2,000 records and 9 features. Hence, the combined dataset includes 2,768 records and 9 features. Every patient in these datasets was older than 21 and female. Combining data from multiple sources can enhance the accuracy of predictive models and increase their ability to generalize across different populations.

B. Initializing and Preprocessing Data

The three datasets have passed through several steps of exploratory process, analysis, and preprocessing before feeding it to the training and testing phase for diagnosing diabetes. These steps can be summarized as follows (the researchers are concerned only with the Pima Indians dataset. Hence, the same procedure will be applied to the other two datasets).

1) Dataset Overview: This step provides a statistical description and data type for every dataset feature. Tables I–III show the statistical description with the data type of dataset features. The Pima Indians dataset in Table I indicates a high risk of diabetes due to higher BMI (31.99 kg/m2) and mean glucose (120.89 mg/dL). This tendency is supported by median results

TABLE II
KEY STATISTICS OF FEATURES IN THE GERMAN SOCIETY DATASET, EMPHASIZING VARIABILITY IN INSULIN LEVELS, REFLECTING DIFFERENCES IN INSULIN RESISTANCE OR TREATMENT STATUS AMONG PARTICIPANTS.

Index	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	Body Mass Index (BMI)	Diabetes Pedigree Function	Age	Outcome
Count	2000	2000	2000	2000	2000	2000	2000	2000	2000
Mean	3.70	121.18	69.15	20.94	80.25	32.19	0.47	33.09	0.34
Std.	3.31	32.07	19.19	16.10	111.18	8.15	0.32	11.79	0.47
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	63.50	0.00	0.00	27.38	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	40.00	32.30	0.38	29.00	0.00
75%	6.00	141.00	80.00	32.00	130.00	36.80	0.62	40.00	1.00
Max	17.00	199.00	122.00	110.00	744.00	80.60	2.42	81.00	1.00
Data Type	int64	int64	int64	int64	int64	float64	float64	int64	int64

TABLE III
COMBINED DATASET STATISTICS, HIGHLIGHTING FEATURE CONSISTENCY ACROSS POPULATIONS, INDICATING SIMILAR HEALTH CHARACTERISTICS ACROSS THE POPULATIONS.

Index	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	Body Mass Index (BMI)	Diabetes Pedigree Function	Age	Outcome
Count	2768	2768	2768	2768	2768	2768	2768	2768	2768
Mean	3.74	121.10	69.13	20.82	80.13	32.14	0.47	33.13	0.34
Std.	3.32	32.04	19.23	16.06	112.30	8.08	0.33	11.78	0.48
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	37.00	32.20	0.38	29.00	0.00
75%	6.00	141.00	80.00	32.00	130.00	36.62	0.62	40.00	1.00
Max	17.00	199.00	122.00	110.00	846.00	80.60	2.42	81.00	1.00
Datatype	int64	int64	int64	int64	int64	float64	float64	int64	int64

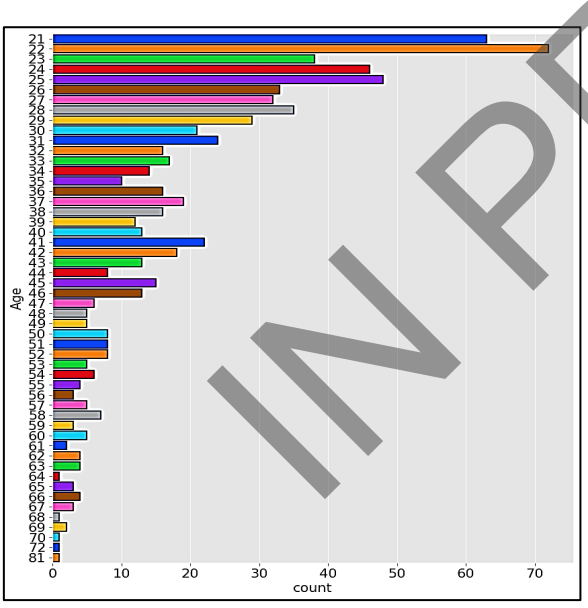

Fig. 1. Frequency distribution of age groups within the Pima Indians dataset.

(glucose: 117; BMI: 32), whereas zeros indicate missing data. Extreme results (e.g., BMI max: 67.1) and high variability (glucose std. dev.: 31.97; BMI: 7.88) highlight metabolic concerns and call for preprocessing for reliable modeling.

Clinically, significant diabetes risk indicators are found in the German Society dataset (see Table II), with mean glucose (121.18 mg/dL) and BMI (32.19 kg/m2) beyond healthy standards. The presence of prediabetes or diabetes is confirmed by the 75th percentile glucose level (141 mg/dL), but extremes in BMI (maximum 80.6) signify severe obesity. Population variability is crucial for predictive modeling, as indicated by standard deviations (glucose: 32.07, BMI: 8.15).

Among the combined dataset in Table III, mean glucose (121.10 mg/dL) and BMI (32.14 kg/m2) are clinically significant, maintaining similar diabetes risk patterns. Important findings include excessive BMI values (max 80.6) suggesting severe obesity, repeated zero-minimum values necessitating data imputation, and 75th percentile glucose (141 mg/dL) confirming diabetes risk. The requirement for significant preprocessing is highlighted by population variability (glucose standard deviation: 32.04, BMI standard deviation: 8.08).

2) Exploratory Data Analysis: Analysis of the datasets is done regarding the two variables of gender and age. The patients are all female and over 21, as already mentioned. The frequencies of ages in the dataset are illustrated in Fig. 1. Young individuals (ages 21 to 40) make up the majority of participants, whereas fewer represent older individuals (ages 81 and

```
<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
```

Fig. 2. The missing values of the Pima Indians dataset.

higher). The rate of diabetes development rises with age, especially beyond age 40, and is consistent with the recognized evolution of metabolic risk.

3) Feature Exploration - Statistic Approach: The outliers and missing values are identified in this step. There are no missing values in the datasets. The elimination resolves the issue with outliers in each feature. The missing values and outliers in the Pima Indians dataset are illustrated in Figs. 2 and 3.

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 about the method used [26]. 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 using the Sequential Backward Selection (SBS) technique. The SBS algorithm is first proposed by Marill and Green (1963). SBS has the advantage of starting with a full set of features and gradually removing the least influential features, which makes it useful when anticipating that some features may be ineffective or redundant. SBS is effective when there is a strong correlation between some features, as it helps to eliminate redundant features while keeping the best features. Dimensionality reduction is known to reduce the complexity of the model and the likelihood of overfitting, which can be effectively achieved using SBS. During the selection process, once the features are removed from the set, they cannot be included again. It causes the methods to give optimal results [27, 28]. The most important features in the three datasets obtained using the SBS algorithm are ['Glucose', 'Body Mass Index (BMI)', 'Age', 'Pregnancies'].

5) Imbalance Data Handling: A class is considered the majority if it contains more observations in a dataset than the other classes. In other words, if the observations in a database for a given class are less than the other classes in the same database, the class is considered to be a minority. Such datasets are called imbalanced datasets [29, 30]. Imbalanced datasets can be encountered in a wide range of practical applications, including medical diagnosis. In the research, the oversampling method using the Synthetic Minority Oversampling Technique (SMOTE) technique is used to address the issue of imbalanced data in the Pima Indians dataset.

Oversampling aims to equalize the class distribution by multiplying minority class data. 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 [31]. The most commonly used oversampling method is the SMOTE approach [32]. Unlike random sampling, this method creates synthetic data by analyzing existing minority data. SMOTE cannot perfectly represent how the original samples are distributed. Byh7in the new synthetic samples. Therefore, the performance of the classifier may be impacted by errors in the distribution of data when utilizing SMOTE-based oversampling techniques. It can increase the probability that the samples will be misclassified [33]. Figure 4 shows the distribution of Pima Indians dataset classes in the original, undersampling, and oversampling datasets.

C. Construction of Models

The Ensemble algorithms are among the most effective tools in the field of machine learning, as they rely on the idea of combining several predictive models to build a stronger and more accurate model. These algorithms rely on combining the strengths of individual models to reduce the weaknesses of each. Some popular algorithms in this field are Gradient Boosting, Random Forest, and AdaBoost. Each of these algorithms uses different approaches to improve predictions by ensemble models, which enhances the performance of individual models and reduces prediction errors.

One of the most common and frequently applied reinforcement learning strategies is Gradient Boosting. The foundation of this algorithm is the gradual construction of a series of weak learners, in which the errors of the previous model are corrected at each

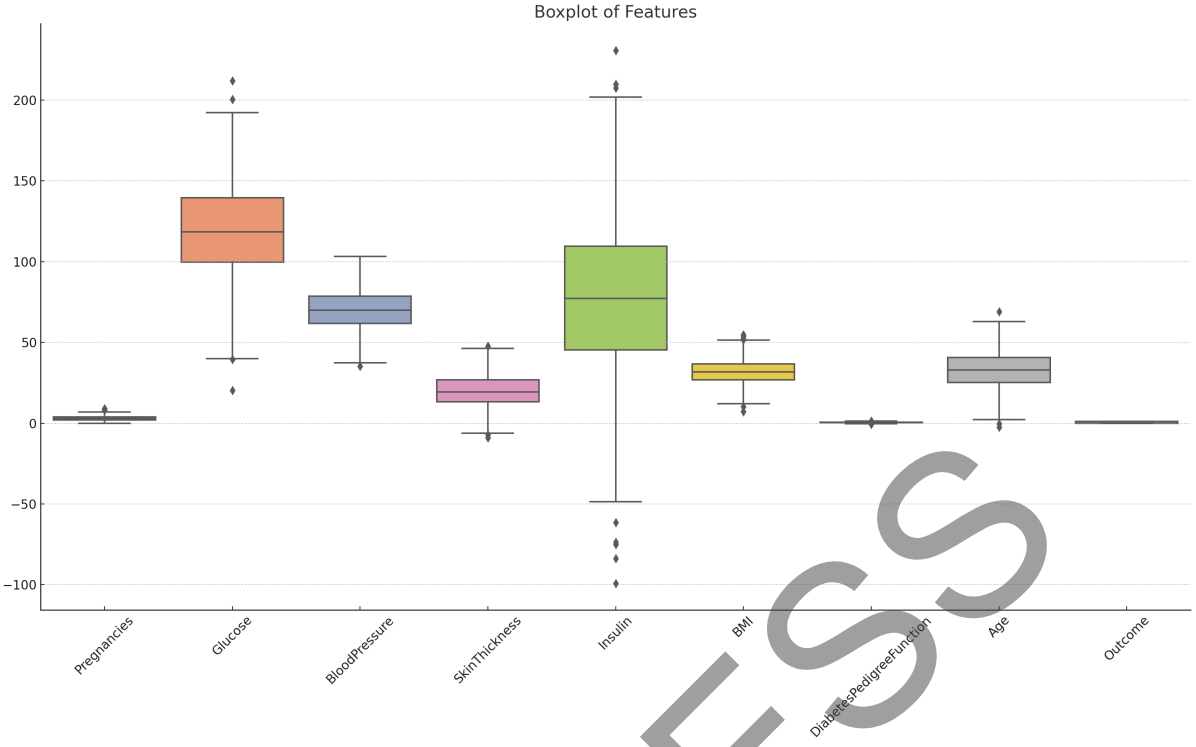

Fig. 3. The outliers values of the Pima Indians dataset.

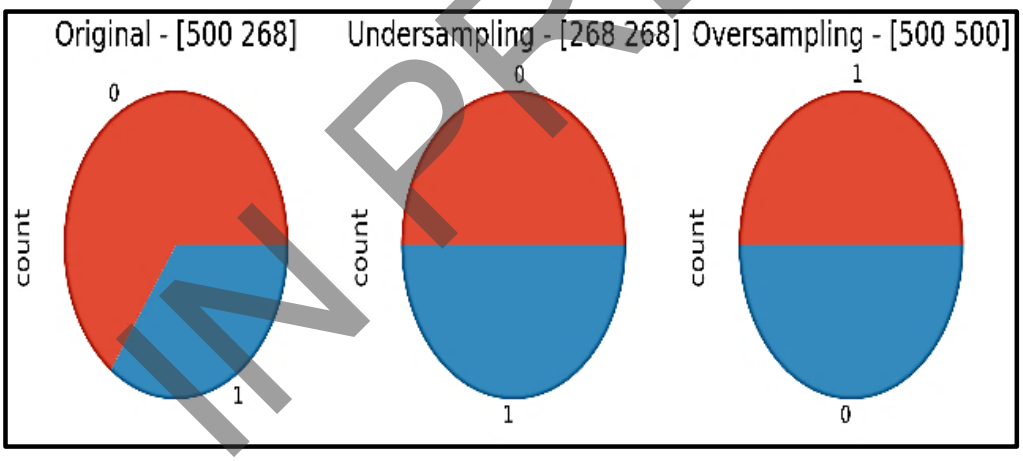

Fig. 4. Original, Undersampling, and Oversampling of Pima datasets.

stage by decreasing the loss function. Typically, this algorithm starts with small Decision Trees as base models, adding new models at each training step to enhance the model’s performance. Next, the residuals produced by this model are computed. Then, using these errors as training data, a new model is created and added to the prior ones. Until a robust model is achieved, this process is repeated multiple times. This algorithm has the advantage of being highly capable of handling high-dimensional data and flexible enough to adapt to various data types. Its drawbacks are that it takes a long time to train, and in an improperly organized model, it is prone to overfitting [34, 35].

One popular ensemble classification technique in machine learning and data science with a broad range of applications is the Random Forest method, first presented by Breiman (2001) [34]. Regression can be made using the supervised machine learning technique

TABLE IV
THE BEST HYPERPARAMETERS TUNING OF ENSEMBLE MACHINE LEARNING ALGORITHMS.

Dataset	Model	Best Parameters
Pima Indians Diabetes Dataset	Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
	Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 100}
	AdaBoost	{'learning_rate': 0.1, 'n_estimators': 200}
German Society Dataset	Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 200}
	Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
	AdaBoost	{'learning_rate': 1.0, 'n_estimators': 200}
Combined Dataset	Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 200}
	Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
	AdaBoost	{'learning_rate': 1.0, 'n_estimators': 200}

known as the Random Forest algorithm, which is categorized based on the nature of the problem. With the help of this technique, multiple Decision Trees are randomly combined, their predictions are collected, and a decision is made based on the majority vote of the Decision Trees. It is important to remember that the Random Forest algorithm performs optimally in environments where the number of variables is much greater than the number of observations [36]. The advantages of this algorithm are that it greatly reduces overfitting and has the ability to handle missing data and high data variability. However, it requires a large amount of memory due to the large number of trees, as well as the difficulty of interpreting the final model.

Adaptive Boosting, or AdaBoost, is a widely used and popular reinforcement learning technique. The foundation of this technique is the improvement of basic models' performance by the use of data that is incorrectly classified in previous iterations. Then, sampling weights are modified to provide more weight to data that are challenging to classify. The approach first uses a weak model (a basic Decision Tree, for example) and modifies the sampling weights to give samples that are incorrectly classified greater weight. This process is performed numerous times to create a robust model. This algorithm has the advantage of being able to effectively adapt to a variety of data types and greatly increase the accuracy of simple models. Its sensitivity to highly noisy data and reliance on selecting the right baseline model are its drawbacks [34, 37].

III. RESULTS AND DISCUSSION

The researchers show the results of using three different diabetes datasets to predict diabetes using a variety of intelligent algorithms. The performance evaluation of several ensemble learning algorithms, such as Gradient Boosting, Random Forest, and AdaBoost, is the main goal of the research. Key assessment measures like F1-score, ROC-AUC, accuracy, precision, recall, log loss, and cross-validation accuracy are used to assess the outcomes. The best model for diabetes prediction is found after a thorough analysis and discussion of the data.

Multiple features linked to medical conditions that may have a role in the development of diabetes are present in the datasets used. Before being used, the dataset is divided into training (80%) and testing sets (20%), with the features standardized. Next, the ensemble machine learning algorithms used in the research are trained and tested, with hyperparameters tuned using the Grid Search method.

Table IV presents the best hyperparameters that give the best performance. The seven measures are selected to offer a thorough and impartial evaluation of the models' effectiveness in diabetes prediction. While precision and recall aim to reduce false positives and false negatives, respectively, accuracy represents the total percentage of correct predictions and ensures that impacted instances are identified. Because it strikes a compromise between recall and precision, the F1 Score may be used with unbalanced data. Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) evaluates the model's capacity to differentiate between those who are impacted and those who are not across various thresholds. Log Loss assesses probability accuracy, which shows how confident the model is in its forecasts. Lastly, cross-validation strengthens the results' generalizability and stability, reinforcing model dependability. The results obtained from these models are summarized and discussed as follows.

The Random Forest model outperforms the other ensemble models in the Pima Indians dataset (see Table V). It achieves an F1-score of 0.827, an accuracy of 0.817, and a precision of 0.785. With slightly worse performance, the Gradient Boosting and AdaBoost models' accuracies range from 0.791 to 0.796. Figure 5 shows the ROC-AUC of ensemble algorithms for the Pima Indians dataset. The Random Forest model outperforms because it combines multiple trees that work independently, which allows it to handle variability in the data efficiently. Since the Pima Indians dataset contains heterogeneous features or noise, the Random Forest's ability to assign different weights to each

TABLE V
THE PERFORMANCE EVALUATION METRICS OF ENSEMBLE LEARNING ALGORITHMS ON PIMA INDIANS DATASET.

	Accuracy	Precision	Recall	F1-Score	Receiver Operating Characteristic - Area Under the Curve (ROC-AUC)	Log Loss	Cross-Validation Accuracy
Random Forest	0.817	0.785	0.875	0.827	0.916	0.379	0.813
Gradient Boosting	0.791	0.764	0.843	0.801	0.915	1.008	0.814
AdaBoost	0.796	0.766	0.854	0.807	0.864	0.648	0.801

TABLE VI
THE PERFORMANCE EVALUATION METRICS OF ENSEMBLE LEARNING ALGORITHMS ON GERMAN SOCIETY DATASET.

	Accuracy	Precision	Recall	F1-Score	Receiver Operating Characteristic - Area Under the Curve (ROC-AUC)	Log Loss	Cross-Validation Accuracy
Gradient Boosting	0.996	0.992	1.000	0.996	0.999	0.035	0.961
Random Forest	0.994	0.992	0.996	0.994	0.998	0.087	0.958
AdaBoost	0.860	0.829	0.908	0.866	0.939	0.677	0.860

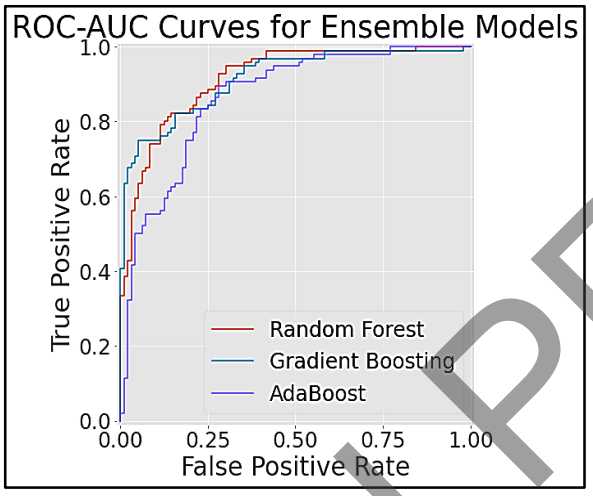

Fig. 5. The Receiver Operating Characteristic (ROC) curves for different ensemble machine learning models applied to the Pima Indians dataset.

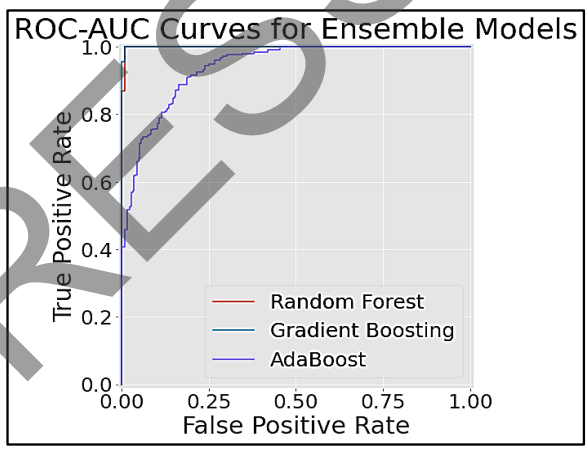

Fig. 6. The Receiver Operating Characteristic (ROC) curves for different ensemble machine learning models applied to the Germany Society dataset.

feature helps it to perform better.

Meanwhile, the models with the highest accuracy on the German Society dataset are Gradient Boosting and Random Forest, with respective accuracies of 0.996 and 0.994. With an accuracy of 0.860, AdaBoost also performs well (see Table VI and Fig. 6). Gradient Boosting works effectively for features that need to be improved gradually since it is based on fixing the errors of previous models. Correcting cumulative errors may make features in the German Society dataset stand out better, improving the effectiveness of Gradient boosting.

Using the combined dataset, Gradient Boosting and Random Forest yield the best accuracy of 0.991 and 0.988, respectively, as shown in Table VII. With an accuracy of 0.837 and an F1-score of 0.833, the AdaBoost model also demonstrates good behavior. Figure 7 illustrates the ROC-AUC of ensemble algorithms for a combined dataset. There can be greater variation in patterns or distributions in data combined from different sources, and this is where Gradient Boosting is appropriate since it learns cumulatively from model errors. This property may make it an ideal model for data that combines different groups, as it builds robust models that adapt to a variety of patterns.

The results show how different machine learning techniques perform differently on different datasets. When it came to balanced performance criteria and generalizability, in particular, ensemble learning strategies perform better than traditional methods overall

TABLE VII
THE PERFORMANCE EVALUATION METRICS OF ENSEMBLE LEARNING ALGORITHMS ON COMBINED DATASET.

	Accuracy	Precision	Recall	F1-Score	Receiver Operating Characteristic - Area Under the Curve (ROC-AUC)	Log Loss	Cross-Validation Accuracy
Gradient Boosting	0.991	1.000	0.982	0.991	1.000	0.027	0.983
Random Forest	0.988	1.000	0.976	0.988	0.999	0.056	0.981
AdaBoost	0.837	0.853	0.813	0.833	0.935	0.680	0.859

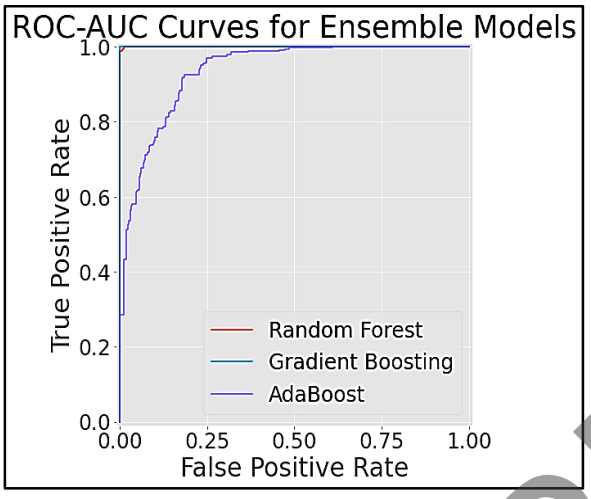

Fig. 7. The Receiver Operating Characteristic (ROC) curves for different ensemble machine learning models applied to a combined dataset.

compared with literature studies. The dataset’s features, however, significantly influence each technique’s level of effectiveness.

The research findings have a significant impact on the development of diabetes prediction models. Using ensemble techniques makes sense for building robust and widely applicable models because they outperform other approaches. However, the variations in dataset performance show how important it is to evaluate and adjust models uniquely for each dataset.

These findings demonstrate how ensemble learning techniques may detect intricate, nonlinear relationships in data, which makes them highly valuable for diabetes prediction applications. Additionally, the results show that merging data from many sources can improve the accuracy and population-generalizability of models. The application of deep learning approaches to these datasets should be the focus of future studies because they have demonstrated promise in a number of predictive modeling domains. Moreover, incorporating more varied datasets can potentially improve the models’ applicability.

IV. CONCLUSION

The research examines the application of ensemble learning methods to predict the onset of diabetes in three distinct datasets: German, Pima Indians, and combined datasets. The primary objective is to develop a robust and consistent predictive model that can reliably predict diabetes in different populations since it will allow the model to fully represent the intricate patterns and variances found in these datasets. Using a methodology that includes handling missing values, eliminating outliers, choosing features, transforming features, controlling imbalances, and normalizing data, the researchers can prepare its data for training. Afterward, training and testing sets of data are created. For best results, hyperparameters are optimized to construct ensemble learning models.

In the research, the effect of merging different databases on diabetes prediction using ensemble learning techniques is investigated. The results show that integrating data from multiple sources can contribute to improving the accuracy of models and their generalizability to different populations. By comparing the performance of the separate databases of the ”Pima” and ”German” databases and the combined database, it can be seen that the models trained on the combined data show improved performance across all metrics. For example, the Gradient Boosting model on the combined database achieves an accuracy of 0.991 and a ROC-AUC of 1.0, outperforming the same model trained on the “Pima” or “German” database alone.

This performance improvement can be explained by the fact that merging provides a more comprehensive model that can accommodate diversity in the data, making it more generalizable when applied to different populations. Thus, the use of combined data is an important step towards developing more robust and accurate predictive models in diabetes diagnosis, enhancing the possibilities of their use in real-world applications across diverse populations. In addition, the results show that using ensemble learning techniques such as Gradient Boosting and Random Forest can be particularly effective when dealing with combined data, as they can handle the complexity and noise

in the data more effectively than using data from a single source. Ultimately, the research highlights the importance of adopting data-integrating strategies to develop more accurate and reliable predictive models, which may contribute to improving the diagnostic process and increasing the chances of early detection of diabetes across different populations.

Future research should concentrate on applying deep learning techniques to these datasets as they have shown promise in a variety of predictive modeling fields. Additionally, adding a wider variety of datasets may enhance the models' applicability. The use of deep learning networks, such as Deep Neural Networks, Convolutional Neural Networks (CNN), or Recurrent Networks (RNN), is an important step as these models provide a high ability to detect complex patterns and nonlinear interaction between features, which may further improve predictive performance. The complexity of deep models also requires future research to be directed towards building interpretable models, using techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to understand the impact of different features on prediction results, allowing researchers and clinical practitioners to rely on these models with greater confidence.

Author Contribution

Conceived and designed the analysis, E. M. H., H. J., and A. A. A. I.; Collected the data, E. M. H., H. J., and A. A. A. I; Contributed data or analysis tools, E. M. H., H. J., and A. A. A. I; Performed the analysis, E. M. H., H. J., and A. A. A. I; Wrote the paper, E. M. H.; and Reviewed the paper, H. J. and A. A. A. I.

Data Availability

The data that support the findings of the research are openly available in the Kaggle repository at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database and https://www.kaggle.com/competitions/diabetes, reference number [24, 25].

References

[1] National Institute of Diabetes and Digestive and Kidney Diseases, "What is diabetes?" 2023. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[2] International Diabetes Federation (IDF), "IDF diabetes atlas 11th edition," 2025. [Online]. Available: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[3] A. Saini, K. Guleria, and S. Sharma, "Predictive machine learning techniques for diabetes detection: An analytical comparison," in *2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)*. Rajpura, India: IEEE, Feb. 24–26, 2023, pp. 1–5.

[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.

[5] A. Yaganteeswarudu, "Multi disease prediction model by using machine learning and Flask API," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. Coimbatore, India: IEEE, June 10–12, 2020, pp. 1242–1246.

[6] E. M. Hameed, I. S. Hussein, H. G. A. Altameemi, and Q. K. Kadhim, "Liver disease detection and prediction using SVM techniques," in *2022 3rd Information Technology to Enhance E-Learning and Other Application (IT-ELA)*. Baghdad, Iraq: IEEE, Dec. 27–28, 2022, pp. 61–66.

[7] R. Alhalaseh, D. A. G. AL-Mashhadany, and M. Abbadi, "The effect of feature selection on diabetes prediction using machine learning," in *2023 IEEE Symposium on Computers and Communications (ISCC)*. Gammarth, Tunisia: IEEE, July 9–12, 2023, pp. 1–7.

[8] E. M. Hameed and H. Joshi, "Performance comparison of machine learning techniques in prediction of diabetes risk," in *AIP Conference Proceedings*, vol. 3051, no. 1. Al-Samawa, Iraq: AIP Publishing, May 3–4, 2024.

[9] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *Journal of Diabetes & Metabolic Disorders*, vol. 23, no. 1, pp. 603–617, 2024.

[10] Q. Zou, Y. Zhang, and C. S. Chen, "Construction and application of a machine learning prediction model based on unbalanced diabetes data fusion," in *Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence*. Shanghai, China: Association for Computing Machinery, July 7–9, 2023, pp. 114–123.

[11] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019*. Sikkim, India: Springer,

Feb. 26–27, 2020, pp. 113–125.

[12] 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, no. 1, pp. 345–365, 2025.

[13] P. Chen and C. Pan, "Diabetes classification model based on boosting algorithms," *BMC Bioinformatics*, vol. 19, pp. 1–9, 2018.

[14] S. Joshi and S. R. PriyankaShetty, "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, no. 3, pp. 1168–1173, 2015.

[15] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[16] P. Cıhan and H. Coşkun, "Performance comparison of machine learning models for diabetes prediction," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. Istanbul, Turkey: IEEE, June 9–11 2021, pp. 1–4.

[17] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on Machine Learning (ML) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16 157–16 173, 2023.

[18] B. Farajollahi, M. Mehmannavaz, H. Mehrjoo, F. Moghbeli, and M. J. Sayadi, "Diabetes diagnosis using machine learning," *Frontiers in Health Informatics*, vol. 10, no. 1, pp. 1–5, 2021.

[19] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big data*, vol. 6, no. 1, pp. 1–19, 2019.

[20] S. Kumari and S. Kumar, "A comparative study of various data transformation techniques in data mining," *International Journal of Scientific Engineering and Technology*, vol. 4, no. 3, pp. 146–148, 2015.

[21] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022.

[22] A. G. Karegowda, V. Punya, M. A. Jayaram, and 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, no. 12, pp. 45–50, 2012.

[23] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data-mining technologies for diabetes: A systematic review," *Journal of Diabetes Science and Technology*, vol. 5, no. 6, pp. 1549–1556, 2011.

[24] John, "Diabetes." [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes

[25] L. Hernandez, "Diabetes," 2019. [Online]. Available: https://kaggle.com/competitions/diabetes

[26] A. A. Alhussan, A. A. Abdelhamid, S. K. Towfek, A. Ibrahim, M. M. Eid, D. S. Khafaga, and M. S. Saraya, "Classification of diabetes using feature selection and hybrid Al-Biruni earth radius and dipper throated optimization," *Diagnostics*, vol. 13, no. 12, pp. 1–40, 2023.

[27] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 1–17, 2022.

[28] M. R. Alnowami, F. A. Abolaban, and E. Taha, "A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer," *Journal of Radiation Research and Applied Sciences*, vol. 15, no. 1, pp. 104–110, 2022.

[29] A. L. Lynam, "Developing clinical prediction models for diabetes classification and progression," Ph.D. dissertation, University of Exeter, 2019.

[30] E. M. Hameed and H. Joshi, "Improving diabetes prediction by selecting optimal K and distance measures in KNN classifier," *Journal of Techniques*, vol. 6, no. 3, pp. 19–25, 2024.

[31] F. Mesquita, J. Maurício, and G. Marques, "Oversampling techniques for diabetes classification: A comparative study," in *2021 International Conference on e-Health and Bioengineering (EHB)*. Iasi, Romania: IEEE, Nov. 18–19, 2021, pp. 1–6.

[32] M. Shuja, S. Mittal, and M. Zaman, "Effective prediction of type II diabetes mellitus using data mining classifiers and SMOTE," in *Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019*. Rajasthan, India: Springer, April 13–14, 2020, pp. 195–211.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[34] H. Ait Naceur, H. G. Abdo, B. Igmoullan, M. Namous, F. Alshehri, and J. A. Albanai, "Implementation of random forest, adaptive boosting, and

gradient boosting decision trees algorithms for gully erosion susceptibility mapping using remote sensing and GIS," *Environmental Earth Sciences*, vol. 83, no. 3, pp. 1–20, 2024.

[35] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[36] R. Natras, B. Soja, and M. Schmidt, "Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting," *Remote Sensing*, vol. 14, no. 15, pp. 1–34, 2022.

[37] R. Kumar, B. Rai, and P. Samui, "A comparative study of AdaBoost and k-nearest neighbor regressors for the prediction of compressive strength of ultra-high performance concrete," in *Recent Developments in Structural Engineering, Volume 1*. Nagpur, India: Springer, Dec. 7–9, 2023, pp. 23–32.