Ensemble IDO Method for Outlier Detection and N_2O Emission Prediction in Agriculture

Ahmad Rofiqul Muslikh, Pulung Nurtantio Andono, Aris Marjuni, Heru Agus Santoso
Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia

***Abstract*—Nitrous oxide (N_2O) emissions from agricultural activities significantly contribute to climate change, necessitating accurate predictive models to inform mitigation strategies. This study proposes an ensemble framework combining Isolation Forest, DBSCAN, and One-Class SVM to enhance outlier detection in N_2O emission datasets. The dataset, consisting of 2,246 rows and 21 columns, was preprocessed to address missing values and normalize data. Outlier detection was performed using each method individually, followed by integration through hard and soft voting techniques. The results revealed that Isolation Forest identified 113 outliers, DBSCAN detected 1,801, and One-Class SVM found 118. Hard voting identified 165 outliers, while soft voting detected 734, ensuring a refined dataset for subsequent modeling. The ensemble approach improved the accuracy of the XGBoost model for N_2O emission prediction. The best results were obtained using the Random Search Cross Validation hyperparameter tuning, with a test size is 20%, achieving a CV MSE of 0.0215, MSE of 0.0144, RMSE of 0.1200, MAE of 0.0723, and an R^2 of 0.6750. This study demonstrates the effectiveness of combining multiple outlier detection methods to enhance data quality and model performance, supporting more reliable predictions of N^2O emissions.**

Keywords*—Ensemble framework; outlier; detection; N2O emission; isolation forest; DBSCAN; one-class SVM

I. Introduction

Nitrous oxide (N2O) emissions from agricultural activities significantly threaten climate stability due to their high global warming potential [1], approximately 298 times greater than carbon dioxide [2]. Accurate prediction of N2O emissions is essential for effective environmental management and climate change mitigation. However, existing predictive models often struggle with outliers, which can skew results and reduce model accuracy [3]. Recent studies have highlighted the complexity of predicting N2O emissions due to various influencing factors, such as soil type, climatic conditions, and agricultural practices [4], [5]. Traditional predictive models, ranging from empirical observational models to more complex process-based models, face significant challenges in handling outliers, resulting from measurement errors, extreme weather events, or anomalies [6]. Effectively identifying and handlinghese outliers are crucial for improving model accuracy and reliability [7].

Outlier detection plays a critical role in enhancing the accuracy of predictive models. Various methods, such as Isolation Forest, DBSCAN, and One-Class SVM, have been effective in identifying outliers in environmental data [6], [3], [8]. These methods are essential for ensuring data quality and improving the reliability of predictive models used for N_2O emission analysis [9]. However, using these methods individually has limitations regarding parameter sensitivity and scalability. The proposed IDO ensemble framework combines these methods to provide a more robust and accurate outlier detection mechanism.

The comparative results differ across datasets due to varying data characteristics such as density, distribution, and noise levels. Isolation Forest an ensemble method isolates observations by randomly selecting a feature and then choosing a split value between the maximum and minimum values of the selected feature [10]. Isolation Forest efficiently handles high-dimensional data but may struggle with clustered anomalies. This technique has proven to be robust for detecting various types of outliers in well-log datasets, achieving an accuracy of 90.2% in distinguishing between inliers and outliers [11], [12].

Its efficiency in handling high-dimensional data makes it suitable for large and complex datasets. DBSCAN, a density-based clustering algorithm, identifies core, borderand noise points based on a specified radius and minimum number of points [13]. DBSCAN excels at identifying clusters in noisy data but requires precise parameter tuning. This method effectively detects noise and manages noisy datamaking it valuable for environmental data analysis where noise is common [11].

One-Class SVM, a machine learning algorithm for anomaly detection, constructs a boundary around normal data points to identify outliers [11]. One-Class SVM effectively defines decision boundaries in complex feature spaces but is sensitive to kernel choices. This technique is outstanding in detecting anomalies with high correctness, distinctiveness, and robustness, proving to be particularly useful in identifying rare but significant anomalies in agricultural datasets [9].

Although Isolation Forest, DBSCAN, and One-Class SVM each have unique strengths, using them individually has parameter sensitivity and scalability limitations. Combining these methods into an ensemble can provide more robust and accurate outlier detection [6] [14]. This ensemble framework leverages the strengths of each algorithm while mitigating their inherent weaknesses. Isolation Forest excels at managing high-dimensional data but can struggle with detecting clustered anomalies. DBSCAN is proficient at identifying clusters and noise but demands meticulous parameter tuning. One-Class SVM effectively defines decision boundaries but is sensitive to kernel choices.

By integrating these methods, this ensemble framework provides a comprehensive outlier detection mechanism, improving data quality and model performance.

The ensemble approach reduces reliance on precise parameter settings for any single method, thereby enhancing overall robustness. The ensemble method efficiently handles high-dimensional data by utilizing the strengths of Isolation Forest and One-Class SVM, while DBSCAN manages dense clusters and noise. The hard and soft voting mechanisms ensure that outliers identified by multiple methods are more likely to be genuine anomalies [15] [18], thereby reducing the likelihood of false positives and false negatives. By effectively identifying and removing outliers, the ensemble framework ensures higher quality data, leading to improved predictive model performance.

By integrating these methods, the IDO framework ensures comprehensive outlier detection, adapting to different data characteristics and improving overall model eprformance. For instance, Isolation Forest's random partitioning effectively isolates anomalies in datasets with high-dimensional features. In contrast, DBSCAN performs better in datasets with dense clusters and noise, identifying core points and noise points. One-Class SVM excels in scenarios with complex decision boundaries, distinguishing normal data from anomalies. The ensemble approach leverages these strengths, ensuring robust outlier detection across various datasets, thereby enhancing the quality and reliability of predictive models.

In addition to outlier detection, this study focuses on predictive modeling of N_2O emissions using the XGBoost algorithm. Known for its high performance and efficiency in handling large datasets, XGBoost has shown superior predictive capabilities compared to traditional models [16], [14]. Hyperparameter tuning is crucial for maximizing model performance, and this study compares the untuned XGBoost model with models optimized through Grid Search and Random Search techniques [17], [18], [19].

The models are evaluated using cross-validation techniques to assess their robustness and generalizability. Cross-validation helps mitigate the risk of overfitting by validating the model on different subsets of the data, ensuring comprehensive and robust performance evaluation [20]. Validation measures play a critical role in ensuring the robustness of predictive models.

Cross-validation techniques, including KFold and standard cross-validation, help mitigate overfitting by validating the model on different data subsets, ensuring comprehensive performance evaluation. Conducting thorough comparisons with existing related work is essential to highlight the advancements and improvements brought by the proposed model. Comparing performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 scores across different models provides valuable insights into the effectiveness of the proposed approach. This study also explores the impact of different test sizes on model performance, ensuring that the findings are applicable across various scenarios in agricultural data analysis.

Despite advancements in predictive modelling and outlier detection, integrating these methods effectively remains challenging. This research addresses this gap by developing and evaluating new outlier detection methods suitable for high-dimensional agricultural data and implementing ensemble methods to enhance robustness [21]. The study aims to improve the predictive accuracy of N_2O emissions models, providing valuable insights for environmental management and contributing to effective climate change mitigation strategies.

II. METHOD

To address the challenges in predicting N2O emissions and to enhance the accuracy of outlier detection, this study employs a comprehensive methodology combining advanced statistical techniques and ensemble learning.

A. Dataset

In this study, we utilized a comprehensive public dataset on nitrous oxide (N2O) emissions from agricultural activities provided by Saha et al. (2021) [21]. The dataset spans from 2002 to 2014, encompassing 2,246 entries and 21 distinct variables. This dataset is instrumental in exploring the effects of agricultural practices and environmental conditions on N2O emissions. It facilitates robust outlier detection and supports reproducibility, enabling comparative analyses across different studies [22][23].

Key variables in the dataset include temporal information such as the date, month, and year of the measurements, experimental details like the type of experiment, the purpose of the data usage, and replication identifiers. Environmental conditions are captured through variables like vegetation type and N_2O concentration, as well as the nitrogen application rate. Additionally, the dataset includes meteorological and soil data, including precipitation levels, air temperature, and days after treatment and seeding. Detailed soil properties such as water-filled pore space at a 25 cm depth, ammonium content, nitrate content, and the proportions of clay, sand, and soil organic matter are also recorded. These variables are crucial for understanding how seasonal conditions and soil characteristics impact N_2O emissions, providing essential insights for developing accurate predictive models and understanding the underlying influencing factors [24][25][26].

B. Data Preprocessing

This study us comprehensive preprocessing techniques, including normalization, data cleaning, and handling missing values, to prepare the N_2O emissions dataset for accurate analysis and effective outlier detection [24]. These steps are crucial for maintaining data quality, ensuring the dataset's suitability for model training, and achieving reliable results in classification and anomaly detection tasks [27]. The inherent null values and outliers in the dataset necessitated thorough preprocessing.

Outlier analysis revealed several types, such as point outliers from potential measurement errors or unusual local conditions, and contextual outliers, which seem normal independently but are abnormal in specific contexts, like unusually low emissions during periods of high microbial activity in winter [28][29]. Moreover, collective outliers can arise when data groups deviaterom the norm due to changes in agricultural practices [30]. Global outliers, representing extreme values beyond the typical data range, indicate rare events not accounted for by existing conditions or strategies [31]. These variations highlight the need for robust detection techniques to manage agricultural data's complexities.

Understanding dataset characteristics is essential before applying methods such as data augmentation or outlier detection. This is exemplified in the classification of rice leaf diseases, where model effectiveness is closely linked to the dataset's attributes [32]. Given the dataset's characteristics and the various types of outliers, we explored the application of ensemble methods for outlier detection. Ensemble methods, which combine predictions from multiple models, are recognized for producing more stable and accurate results. Techniques such as Isolation Forest, DBSCAN, and One-Class SVM have proven effective in identifying outliers [9] [11]. These methods complement each other by handling different aspects of outlier detection, such as identifying isolated points or anomalies within dense clusters. By integrating the results from multiple models, ensemble methods enhance the stability and accuracy of predictions, making them particularly suitable for the nuanced analysis required for N_2O emissions in agriculture [33]. This approach ensures a more robust analysis and improves the dataset's quality, facilitating more accurate and reliable N_2O emission predictions.

C. Outlier Detection

Outlier detection is essential for maintaining the accuracy of predictive models in N_2O emissions studies [34]. Outliers, which can result from measurement errors, data entry mistakes, or rare occurrences, significantly affect data analysis if not properly managed. Traditional detection methods, such as statistical tests, visualization, and distance measures, vary in their ability to identify global or local anomalies [35]. This study applies advanced techniques—Isolation Forest (IF), DBSCAN, and One-Class SVM—independently to robustly identify outliers in the N2O emission dataset. The IDO framework integrates these methods, leveraging their strengths to comprehensively address global and local outliers. This multi-method approach enhances the dataset's integrity and significantly improves the performance and reliability of predictive models, highlighting the importance of meticulous data handling in high-quality research.

D. Proposed Ensemble Method

This study introduces an advanced framework called IDO (Isolation Forest, DBSCAN, and One-Class SVM) to improve the detection of outliers in N_2O emission datasets from agricultural activities. The IDO framework combines three established outlier detection techniques into an ensemble approach, enhancing the accuracy and effectiveness of anomaly identification.

Isolation Forest (IF) effectively detect outliers, particularly in high-dimensional datasets. It works by isolating data points using random partitioning, identifying anomalies based on how quickly they can be isolated from the rest of the data [35]. Point x's isolation is measured by the path length h(x), which represents the number of splits required to isolate the point. For a dataset X, the Isolation Forest algorithm can be mathematically described by the following steps:

The first step is Feature Selection, and Random Split process described in (1) outlines the method of randomly selecting a feature (f_j) from the set of all features ($\{f_1, f_2, \dots, f_d\}$) and choose a random split value (s) within the range of this feature. After choosing the feature, a random split value (s) is selected within the range of the chosen feature. This random selection is fundamental to the Isolation Forest algorithm's ability to partition data and isolate anomalies effectively.

$$f_j \in \{f_1, f_2, \dots, f_d\} \text{ and } s \in \left[\min(f_j), \max(f_j)\right] \quad (1)$$

here f_j is a random chosen feature, and s is the split value within the range of f_j.

The second step, Recursive Partitioning process noted on (2), describes a critical step in the Isolation Forest algorithm.

$$\text{Left Child: } \{x \in X \mid x_{f_j} \leq s\}$$

$$\text{Right Child: } \{x \in X \mid x_{f_j} > s\} \quad (2)$$

This involves recursively applying the partitioning to the dataset, which creates a tree structure. The data is split into two subsets based on the selected feature (f_j) and split value (s). This recursive partitioning continues until each data point is isolated in a unique partition, or a predefined maximum tree depth is reached. This iterative splitting is crucial for the algorithm's ability to effectively isolate anomalies within the dataset.

The next step is Tree Construction. This involves recursively continuing the partition until each data point is isolated in its unique partition or the tree reaches a predefined maximum depth.

$$h(x) = \text{number of splits to isolate } x \quad (3)$$

Eq. (3) defines the Path Length $h(x)$ for each data point (x). This path length represents the number of splits or edges traversed from the root of the tree to the point's leaf node. The shorter the path length, the quicker the data point is isolated, indicating it is likely an anomaly. Calculating $h(x)$ is essential for determining how well each data point is isolated within the tree structure.

Eq. (4) describes the Anomaly Scoring process used in the Isolation Forest algorithm. This step utilizes the average path length $h(x)$ to compute the anomaly score for each data point (x). Points with shorter path lengths are more likely to be outliers because random partitions isolate them more quickly. The anomaly score for a data point (x) is given by:

$$\text{Score}(x) = 2^{-\frac{h(x)}{c(n)}} \quad (4)$$

where: $h(x)$ is the average path length of the data point(x), $c(n)$ is a normalization factor approximated by Eq. (5)

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (5)$$

and $H(i)$ is the i-th harmonic number defined as (6)

$$H(i) = \sum_{k=1}^{i} \frac{1}{k} \quad (6)$$

The final step in the Isolation Forest algorithm is Outlier Identification, where data points are classified based on their anomaly scores. A threshold is set to distinguish between normal points and outliers: points with scores above the threshold are considered normal, while those below are deemed outliers.

Following this, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Unlike Isolation Forest, which relies on random partitioning, DBSCAN is a density-based algorithm that clusters data points and identifies outliers as those that do not fit into any cluster [13]. Its effectiveness in identifying clusters and noise in spatial data makes it an apt choice for this ensemble [36]. Since there is no single mathematical equation that defines it, but rather a set of rules describing the clustering process, the general steps of the DBSCAN algorithm are as follows [37] [36].

Eq. (7) define the selecttion a point *P* from the dataset *D* that has not been visited.

$$P \in D \setminus V \tag{7}$$

where *V* is the set of visited points.

Eq. (8) define ϵ-neighborhood $N_\epsilon(P)$ of point *P*, which includes all points within distance ϵ from *P*

$$N_\epsilon(P) = \{Q \in D \mid \text{dist}(P, Q) \leq \epsilon\} \tag{8}$$

where dist (P,Q) is the distance between points P and Q.

Eq. (9) define core point, if the ϵ -neighborhood $N_\epsilon(P)$ contains at least MinPts points, then P is a core point.

$$|N_\epsilon(P)| \geq \text{MinPts} \tag{9}$$

where $|N_\epsilon(P)|$ denotes the cardinality of the ϵ -neighborhood of *P*.

Eq. (10) define cluster formation, if *P* is a core point, then all points *Q* in its ϵ-neighborhood $N_\epsilon(P)$ are added to the same cluster *C*.

$$Q \in N_\epsilon(P) \Rightarrow Q \in C \tag{10}$$

If *P* is associated with multiple clusters, those clusters are merged.

Eq. (11) and Eq. (12) define border point and noise identification. Points *Q* that are within the ϵ-neighborhood of a core point but do not satisfy the *MinPts* condition are classified as border points. Points that are not in the ϵ-neighborhood of any core point are considered noise or outliers.

$Q \in N_\epsilon(P)$ and $|N_\epsilon(Q)| <$ MinPts → Q is border point (11)

$Q \notin N_\epsilon(P)$ for any core point *P* → Q is noise (12)

Eq. (13) define process iteration that repeat the steps until all points in the dataset *D* have been visited.

$$\forall P \in D,\ P \in V \tag{13}$$

Having outlined the DBSCAN algorithm, which excels in identifying clusters and outliers based on density, we now shift our focus to One-Class SVM. This method adopts a different approach, leveraging machine learning techniques to distinguish between normal and anomalous data points. One-Class SVM is particularly useful in scenarios where the dataset contains complex feature spaces, making it a robust choice for detecting outliers in agricultural N2O emission data.

One-Class SVM (Support Vector Machine) is a machine learning technique that models decision boundaries to separate normal data from outliers. It is adept at handling agricultural data, defining the regions in the feature space that correspond to typical data points, thus identifying anomalies outside these regions [38][39]. The following steps outline the One-Class SVM algorithm.

The algorithm starts by defining the One-Class SVM model using a training dataset ($\{x_1, x_2, \ldots, x_N\}$) where each data point x_i belongs to a d-dimensional feature space ($\in \mathrm{R}^{\mathrm{d}}$). Then the kernel selection, as defined by Eq. (14), involves choosing an appropriate kernel function to transform the input data into a higher-dimensional feature space, if needed. This step is crucial for capturing the complex relationships in the data. A commonly used kernel is the Radial Basis Function (RBF) kernel, which is expressed as follows:

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \exp\left(-\gamma |\mathrm{x_i} - \mathrm{x_j}|^2\right) \tag{14}$$

where ($\gamma > 0$) is a parameter that defines the width of the kernel.

Eq. (15), Eq. (16) defines the optimization problem that aims to determine the decision boundary separating the majority of the data points from the outliers. This optimization process identifies the boundary that encloses the normal data within a specified region of the feature space while isolating the anomalies outside this region.

$$\min_{\mathrm{w},\xi,\rho} \frac{1}{2}|\mathrm{w}|^2 + \frac{1}{\nu \mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \xi_\mathrm{i} - \rho \text{ subject to:} \tag{15}$$

$$\left(\mathrm{w} \cdot \phi(\mathrm{x_i})\right) \geq \rho - \xi_\mathrm{i}, \quad \xi_\mathrm{i} \geq 0, \quad \mathrm{i} = 1, \ldots, \mathrm{N} \tag{16}$$

where (w)is the normal vector to the decision boundary, $(\phi(\mathrm{x_i}))$ is the feature mapping, (ξ_i) are slack variables allowing for some margin violations, (ρ) is the offset, and ($\nu \in (0,1]$) controls the fraction of outliers and support vectors.

Eq. (17) define decision function for determining whether a new data point x is an outlier is given by:

$$f(x) = \left(w \cdot \phi(x)\right) - \rho \tag{17}$$

where, A data point x is classified as normal if $(f(x) \geq 0)$ and as an outlier if $(f(x) < 0)$

After defining the optimization problem in Eq. (18), the next step is to identify the support vectors, which are the data points closest to the decision boundary. These vectors are crucial as they shape the boundary and define the margin. The decision function is then applied to classify new data points as normal or outliers based on their position relative to this boundary, effectively separating typical data from anomalies.

Integrating these methods into the IDO ensemble framework leverages the strengths of Isolation Forest, DBSCAN, and One-Class SVM while compensating for their limitations. This combination provides a robust and comprehensive approach to outlier detection, which is essential for analyzing N2O emissions in agriculture, where data precision is critical for developing effective mitigation strategies.

The proposed method in Fig. 1 outlines an innovative ensemble approach to detect outliers in N_2O emission data. This approach combines Isolation Forest, DBSCAN, and One-Class SVM (the IDO model) into an ensemble framework to enhance

accuracy and reliability in identifying outliers in agricultural N_2O emission datasets. The primary objective is to improve the quality of N_2O emission data [11], which will enhance the accuracy of emission prediction models and support efforts to mitigate climate change and promote sustainable agricultural practices.

As shown in Fig. 1, the IDO model framework integrates Isolation Forest, DBSCAN, and One-Class SVM to comprehensively detect outliers using an ensemble method. The process begins with raw data and includes preprocessing steps for outlier detection and result integration. Isolation Forest uses decision trees to isolate outliers, DBSCAN identifies clusters and outliers based on data density, and One-Class SVM uses a hyperplane for differentiation. These methods' results are combined to produce normalized scores, followed by a voting mechanism to identify outliers and set decision boundaries, refining the training dataset. This integrated approach enhances data analysis reliability and accuracy, making it particularly useful for complex environmental and agricultural datasets.

Once outliers are identified and handled, the dataset is split into test data for model evaluation and train data for model training. Feature engineering follows, selecting, transforming, and creating new features from the cleaned training data to optimize the dataset for training. The model is trained on this engineered data, including tuning to enhance performance. After training, cross-validation and performance evaluation validate the model's effectiveness. A validated model confirms its ability to generalize well to new data. The validated model then predicts N2O emissions using test data. This phase evaluates the model against real-world data. Hyperparameter tuning further refines the model parameters, improving accuracy and efficiency. This iterative process creates a feedback loop between feature engineering and parameter optimization.

Hyperparameter optimization is crucial for maximizing model performance [40]. It involves adjusting parameters significantly affecting the model's accuracy and generalization ability [41]. Each algorithm in the ensemble has specific parameters to tune: Isolation Forest adjusts the number of trees and sample size [42], DBSCAN optimizes epsilon and minPts [43], and One-Class SVM tunes nu and gamma for decision margin and complexity [44]. Techniques like Bayesian optimization can efficiently determine the best configurations by modeling performance and selecting the next parameters to test [41].

Proper hyperparameter optimization enhances model accuracy by balancing bias and variance, preventing overfitting and underfitting [40]. Optimized models fully utilize the dataset, providing precise insights for predicting N_2O emissions in agriculture. Integrating these methods within an ensemble framework creates a robust system for outlier detection in high-dimensional environmental data [41]. This approach improves data accuracy, enhancing analysis quality and prediction reliability. Implementing this method increases N_2O emission prediction accuracy, supporting climate change mitigation and sustainable agriculture. This framework combines outlier detection, model tuning, and evaluation into a robust process, ensuring accurate anomaly detection crucial for N_2O emission prediction.

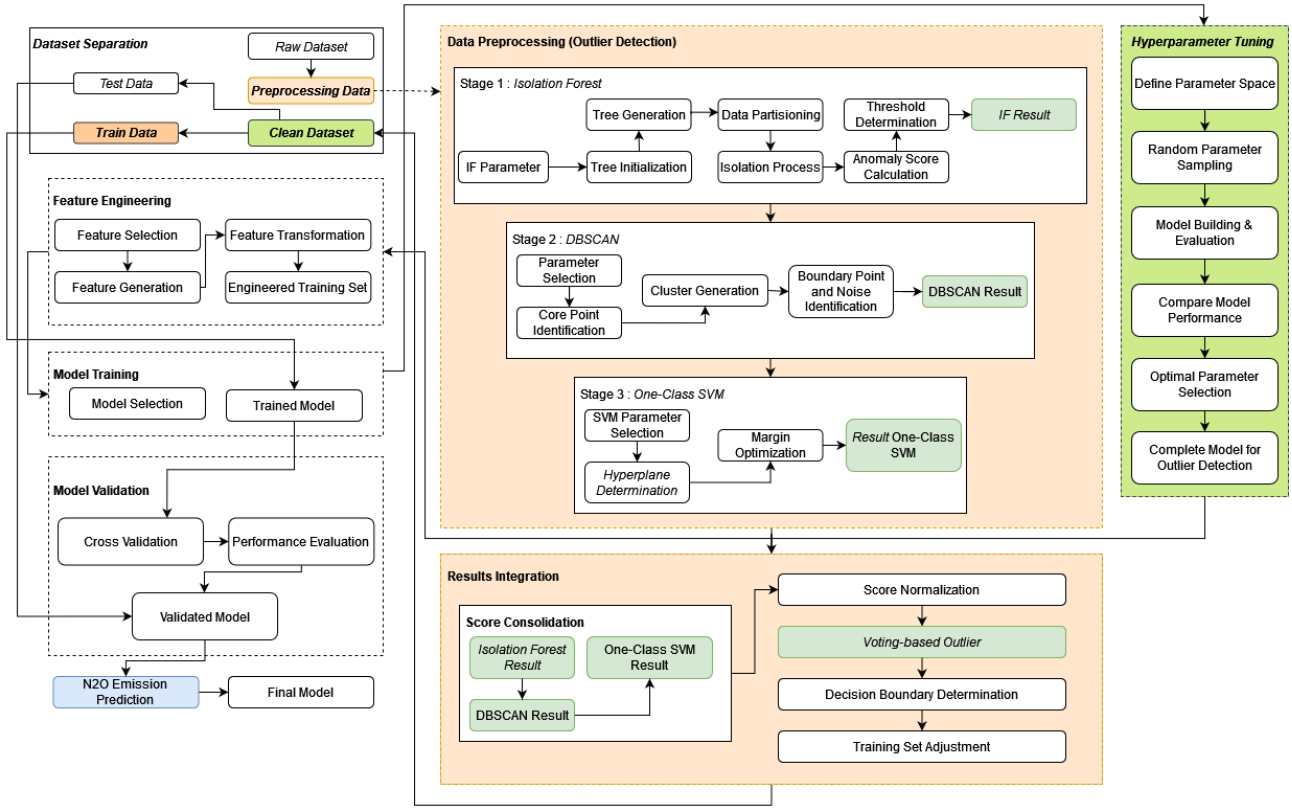

Fig. 1. Proposed model outlier detection with IDO (Isolation Forest – DBSCAN – One-Class SVM) ensemble algorithm.

III. Results

Predicting nitrous oxide (N_2O) emissions in agriculture is challenging due to complex factors. This section examines how advanced machine learning techniques, like outlier detection and ensemble methods, improve the accuracy of N_2O predictions using the XGBoost model.

A. Outlier Detection using Ensemble IDO (IF-DBSCAN-OneClassSVM)

Outlier detection improves model performance by identifying and removing anomalies. Combining Isolation Forest (IF), DBSCAN, and One-Class SVM, the IDO ensemble approach enhanced N_2O emission prediction accuracy. Fig. 2 shows a box plot of N_2O levels, focusing on the 2,072 inliers identified by the IDO method. This cleaned dataset provides a clearer view of the central distribution. The median N_2O level is 1.81, while the mean is 3.22, indicating a slight right skew due to higher inlier values.

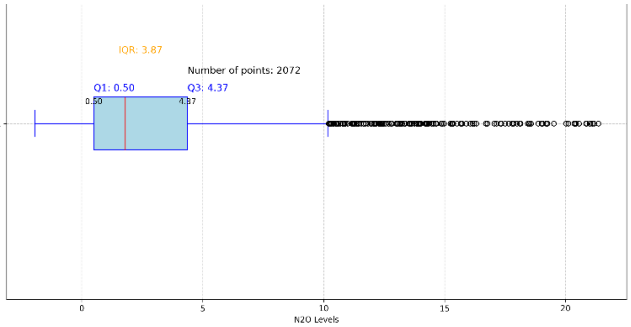

Fig. 2. The inliers identified after applying IDO.

The interquartile range (IQR) for N_2O levels is 3.87, spanning from the 25th percentile (0.50) to the 75th percentile (4.37), showing variability within the central 50% of the data. The range of inliers stretches from -1.94 to 21.36. This indicates that most N_2O levels are concentrated at the lower end but vary significantly within the inliers. Detecting outliers is crucial as it helps exclude extreme values that could distort the dataset. Using the IDO ensemble method effectively removes these outliers, providing a cleaner, more accurate dataset for analyzing and predicting N_2O levels.

Table I summarizes the outlier detection results using Isolation Forest, DBSCAN, and One-Class SVM, highlighting the value of employing multiple techniques. Each method identified distinct sets of outliers, reflecting their unique strengths. Isolation Forest, which isolates points requiring fewer partitions, identified 113 outliers and 2133 inliers. DBSCAN detected 1801 outliers out of 2246 data points, leaving 445 inliers. One-Class SVM, found 118 outliers and 2128 inliers.

TABLE I. N_2O Outlier Detection Result

Method	Outliers	Inliers
Isolation Forest	113	2133
DBSCAN	1801	445
One-Class SVM	118	2128

Applying hard and soft voting methods refined these results, removing the most consistently identified outliers. This process enhanced the dataset's quality and representativeness, crucial for effective predictive modeling.

B. Voting-based Outlier Detection

Integration of outlier detection results from Isolation Forest, DBSCAN, and One-Class SVM was done using Hard and Soft Voting techniques as shown in Table II. The analysis of voting-based outlier detection methods reveals distinct differences in their ability to identify outliers and inliers within the dataset.

TABLE II. Voting-based Outlier Detection Result

Voting Method	Outliers	Inliers
Hard Voting	165	2081
Soft Voting	734	1512

Hard voting identified 165 outliers, leaving 2081 data points as inliers. In contrast, soft voting detected a significantly higher number of outliers, amounting to 734, with the remaining 1512 data points classified as inliers. These results illustrate the varying sensitivity and specificity of the two voting methods, with soft voting being more inclusive in its outlier detection compared to the more stringent hard voting approach.

Post-voting, the XGBoost model's performance was evaluated without cross-validation to set a baseline. Evaluations across test sizes of 20%, 25%, 30%, and 35% aimed to assess the model's robustness and accuracy. Key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score were used to gauge initial effectiveness.

C. XGBoost Model Evaluation with Cross-Validation

In evaluating the XGBoost model's performance, two cross-validation methods were compared: KFold XGBoost (xgb.cv) and Standard Cross Validation (cross_val_score). The results in Tables III and IV, demonstrate the effectiveness of both approaches in enhancing model robustness.

TABLE III. Prediction Evaluation with KFold XGBoost

Test Size	CV MSE	MSE	RMSE	MAE	R2
20%	0.1892	0.0361	0.1900	0.1135	0.1847
25%	0.1917	0.0346	0.1860	0.1139	0.1815
30%	0.1981	0.0330	0.1816	0.1138	0.1848
35%	0.1985	0.0297	0.1724	0.1102	0.1943

KFold XGBoost produced Mean Squared Error (MSE) values ranging from 0.0297 to 0.0361, Root Mean Squared Error (RMSE) values from 0.1724 to 0.1900, Mean Absolute Error (MAE) values from 0.1102 to 0.1139, and R^2 scores between 0.1815 and 0.1943.

TABLE IV. Prediction Evaluation with Standard Cross Validation

Test Size	CV MSE	MSE	RMSE	MAE	R^2
20%	0.0259	0.0143	0.1195	0.0738	0.6776
25%	0.0266	0.0165	0.1286	0.0829	0.6091
30%	0.0269	0.0155	0.1245	0.0791	0.6174
35%	0.0289	0.0152	0.1234	0.0789	0.5876

The results are summarized in Table IV, highlighting various performance metrics. For a test size of 20%, the model exhibited the best performance, achieving the lowest CV MSE of 0.0259 and MSE of 0.0143. Additionally, this configuration resulted in the lowest RMSE of 0.1195 and MAE of 0.0738, along with the highest R^2 score of 0.6776, indicating that the model could explain a substantial portion of the variance in the data.

These results collectively demonstrate that the model performs optimally at a test size of 20%, balancing error metrics and explanatory power. This optimal performance highlights the model's robustness in predicting N_2O emissions under this specific configuration.

D. *Hyperparameter Tuning using GridSearchCV and RandomizedSearchCV*

XGBoost model using two hyperparameter tuning methods: GridSearchCV and RandomizedSearchCV. The results, detailed in Tables V and VI, demonstrate that GridSearchCV slightly outperforms RandomizedSearchCV in terms of key performance metrics but at a higher computational cost.

TABLE V. XGBOOST GRIDSEARCHCV HYPERPARAMETER TUNING EVALUATION

Test Size	CV MSE	MSE	RMSE	MAE	R^2
20%	0.0223	0.0150	0.1223	0.0750	0.6621
25%	0.0230	0.0157	0.1252	0.0807	0.6293
30%	0.0238	0.0151	0.1229	0.0791	0.6268
35%	0.0265	0.0138	0.1175	0.0755	0.6256

GridSearchCV showed Mean Squared Error (MSE) improvements ranging from 2.8% to 6.5%, Root Mean Squared Error (RMSE) improvements from 3.8% to 6.1%, and Mean Absolute Error (MAE) reductions from 4.9% to 10.8% over the untuned model.

TABLE VI. XGBOOST RANDOMIZEDSEARCHCV HYPERPARAMETER TUNING EVALUATION

Test Size	CV MSE	MSE	RMSE	MAE	R^2
20%	0.0215	0.0144	0.1200	0.0723	0.6750
25%	0.0222	0.0152	0.1234	0.0789	0.6397
30%	0.0228	0.0150	0.1224	0.0775	0.6299
35%	0.0255	0.0138	0.1173	0.0756	0.6271

RandomizedSearchCV also improved performance with MSE enhancements from 2.8% to 4.9%, RMSE improvements from 3.9% to 5.7%, and MAE reductions from 4.9% to 8.4%. Although GridSearchCV provided slightly better results, it required significantly more computational resources and time, whereas RandomizedSearchCV was 20-30% faster and more efficient.

Fig. 3 and Fig. 4 visually compare these tuning methods across different test sizes (20%, 25%, 30%, and 35%), focusing on metrics such as Cross-Validation MSE, MSE, RMSE, and MAE. Fig. 3 illustrates the comparison of MSE across different tuning methods. GridSearchCV and RandomizedSearchCV exhibit lower and more stable MSE values than KFoldXGBoost, with StandardCrossVal consistently showing the lowest and most stable MSE values across all test sizes.

In comparison, in Table IV Standard Cross Validation showed significantly better performance with lower MSE values (0.0143 to 0.0165), RMSE values (0.1195 to 0.1286), MAE values (0.0738 to 0.0829), and much higher R^2 scores (0.5876 to 0.6776). as R^2 scores that explain a substantially higher percentage of variance in the data. This suggests that Standard cross-validation provides more reliable and accurate assessments for model evaluation.

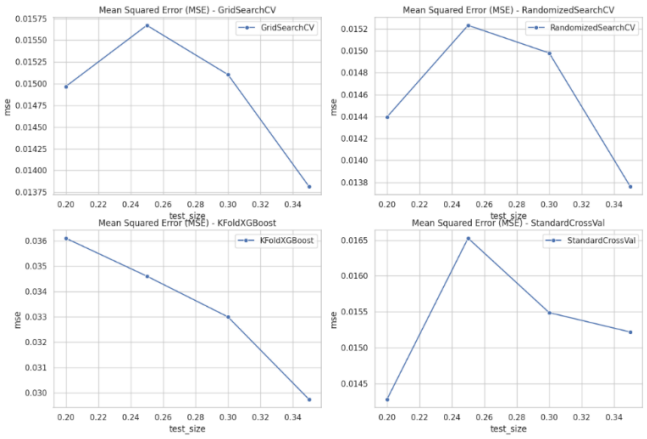

Fig. 3. Comparison of MSE evaluation performance.

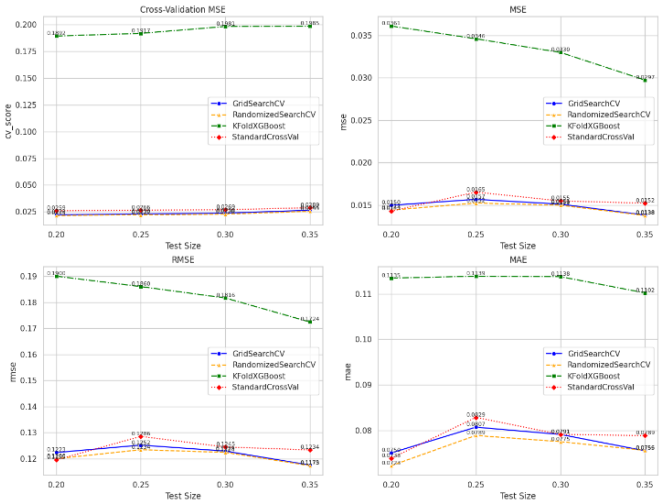

Fig. 4. Comparison of MSE evaluation with different test size.

Fig. 4 expands on these findings by comparing Cross-Validation MSE, RMSE, and MAE across the tuning methods. KFoldXGBoost has the highest Cross-Validation MSE values, indicating higher variance. In contrast, StandardCrossVal and Randomized-SearchCV consistently achieve lower MSE and RMSE values, with StandardCrossVal performing best overall. For example, at a test size of 20%, StandardCrossVal achieves an MSE of 0.0143 compared to KFoldXGBoost's 0.0361, indicating a significant performance advantage.

Similarly, StandardCrossVal shows the lowest RMSE and MAE values, signifying the smallest average prediction errors. At the same test size of 20%, StandardCrossVal achieves an RMSE of 0.1195 and an MAE of 0.0738, compared to KFoldXGBoost's RMSE of 0.1900 and MAE of 0.1135,

showcasing a notable reduction in error rates. Overall, the comparative analysis highlights a trade-off between computational efficiency and model performance.

While GridSearchCV offers marginally better performance, RandomizedSearchCV provides a more practical balance of speed and efficiency, making it suitable for scenarios demanding quicker turnaround times. StandardCrossVal emerges as the most consistent method across all performance metrics, suggesting its robustness and reliability for hyperparameter tuning in XGBoost models.

This analysis emphasizes that the choice of hyperparameter tuning method should consider both the performance improvements and computational resources available. RandomizedSearchCV is an efficient choice for most practical applications, especially in the context of agricultural N_2O emission predictions.

IV. Discussion

In this study presents an advanced ensemble framework that enhances outlier detection in agricultural datasets by combining Isolation Forest, DBSCAN, and One-Class SVM. This integrated approach overcomes issues like parameter sensitivity and scalability associated with individual methods. The researchers processed a dataset containing 2,246 entries and 21 variables, was carefully preprocessed to handle missing values and normalize data.

The ensemble framework's hard and soft voting mechanisms refined outlier detection, identifying 165 outliers with hard voting and 734 with soft voting. Optimized using Random Search Cross Validation, the XGBoost model showed improved predictive performance, with a Mean Squared Error (MSE) of 0.0144 and an R^2 of 0.6750, highlighting the approach's effectiveness in enhancing data quality and supporting accurate N_2O emission predictions critical for climate change mitigation.

Given these findings, it is crucial to delve deeper into the methodologies and their implications on model performance. The study further explores the hyperparameter tuning methods, specifically GridSearchCV and RandomizedSearchCV, and analyzes their impact on the XGBoost model's performance.

A. Comparative Analysis

Outlier detection in N2O emission datasets is challenging due to agricultural data's complexity and high dimensionality. The proposed IDO (Isolation Forest – DBSCAN – One-Class SVM) ensemble framework addresses this by combining three powerful algorithms: Isolation Forest, DBSCAN, and One-Class SVM. Isolation Forest effectively identifies anomalies in high-dimensional data, DBSCAN excels at detecting clusters and differentiating noise based on density, and One-Class SVM distinguishes between normal data and anomalies.

By leveraging these methods through a voting mechanism, the IDO framework ensures accurate outlier detection and enhances data quality, making it highly suitable for analyzing agricultural N2O emissions.

In contrast, previous research has explored different approaches to enhancing outlier detection and clustering validity. For instance, [42] focused on improving cluster validity indices using an ensemble of K-means, K-means++, and Fuzzy C-means clustering algorithms. While this method effectively improved cluster separation, it struggled with robustness in high-dimensional and variable datasets. Similarly, [43] aimed to balance diversity and accuracy in unsupervised outlier ensembles primarily targeting clustering methods. However, this approach often overlooked anomalies that did not form distinct clusters.

Other studies, such as [44], combined multiple detection algorithms to achieve high accuracy in high-dimensional data, but this came at the cost of increased computational complexity. For instance, [45] utilized Isolation Forest with satellite data to detect crop anomalies, achieving high true positive rates but with limited applicability to other types of data. The EBOD method in study [46] effectively handled noisy datasets but lacked the adaptability needed to address the specific challenges of agricultural N2O emissions. While effective in certain contexts, these methods often failed to provide a comprehensive solution for diverse and high-dimensional datasets typical in environmental studies. Table VII shows the comparative analysis of outlier detection methods.

TABLE VII. Comparative Analysis of Outlier Detection Methods

Study	Dataset	Method	Evaluation Metrics	Key Findings
[42]	General clustering datasets	Ensemble of K-means, K-means++, and Fuzzy C-means	Cluster Validity Indices	Improved cluster validity indices post outlier removal. Enhances data quality in various datasets.
[43]	Various real-world datasets	Diversity-Accuracy Balanced Ensemble	True Positive Rate, Diversity-Accuracy Balance	Achieved high true positive rates and balanced detection diversity and accuracy.
[44]	High-dimensional datasets	Ensemble of LOF, KNN, HBOS, iForest, COPOD, PCA	Accuracy, ROC	High accuracy and ROC in detecting outliers in high-dimensional data.
[45]	Sentinel-1 & Sentinel-2 crop data	Isolation Forest with Sentinel-1 and Sentinel-2 data	True Positive Rate	Detected crop anomalies with 94.1% true positive rate for rapeseed and 95.5% for wheat.
[46]	Noisy datasets	EBOD (Ensemble-Based Outlier Detection)	Outlier Detection Accuracy, Noise Robustness	Effective in noisy environments, providing robust outlier detection across various noisy datasets.

Integrating Isolation Forest, DBSCAN, and One-Class SVM, the IDO framework demonstrated superior performance in detecting outliers within N_2O emission datasets. This method identified 113 outliers with Isolation Forest, 1801 with DBSCAN, and 118 with One-Class SVM. Through hard voting, 165 outliers were confirmed, and soft voting identified 734 outliers. This comprehensive detection approach significantly enhanced the dataset's quality and improved the predictive accuracy of the XGBoost model, achieving an R^2 of 0.6750, MSE of 0.0144, RMSE of 0.1200, and MAE of 0.0723. Compared to other methods, the IDO framework provided a more robust, adaptable, and accurate approach for high-dimensional anomaly detection, demonstrating its effectiveness in enhancing N_2O emission predictions.

V. Conclusion

This study highlights the effectiveness of advanced machine learning techniques, particularly cross-validation and hyperparameter tuning, in enhancing the predictive accuracy of the XGBoost model for N2O emissions. Standard Cross Validation outperforms other methods, achieving the lowest errors and highest stability, with significant reductions in RMSE, MAE, and MSE values as low as 0.0143. GridSearchCV delivers slightly better performance metrics but at a higher computational cost, while RandomizedSearchCV provides an efficient alternative with comparable performance improvements. These findings are crucial for improving N2O emission predictions, which are vital for environmental management and climate change mitigation.

Future research should explore sophisticated models and methods, such as deep learning, diverse ensemble learning models, and advanced hyperparameter optimization techniques like Bayesian optimization, to further enhance predictive accuracy and efficiency. Additionally, developing hybrid models and leveraging transfer learning from related datasets could more effectively capture the complex relationships in N2O emissions data. In summary, the choice between GridSearchCV and RandomizedSearchCV depends on balancing computational efficiency and model performance, with RandomizedSearchCV offering a practical solution under computational constraints.

References

[1] S. M. Ogle, K. Butterbach-Bahl, L. Cardenas, U. Skiba, and C. Scheer, "From research to policy: optimizing the design of a national monitoring system to mitigate soil nitrous oxide emissions," Dec. 01, 2020, Elsevier B.V. doi: 10.1016/j.cosust.2020.06.003.

[2] T. T. Nguyen, T. A. T. Pham, and H. T. X. Tram, "Role of information and communication technologies and innovation in driving carbon emissions and economic growth in selected G-20 countries☆," J Environ Manage, vol. 261, May 2020, doi: 10.1016/j.jenvman.2020.110162.

[3] C. Wang, B. Amon, K. Schulz, and B. Mehdi, "Factors that influence nitrous oxide emissions from agricultural soils as well as their representation in simulation models: A review," Apr. 01, 2021, MDPI AG. doi: 10.3390/agronomy11040770.

[4] R. M. Rees et al., "Nitrous oxide emissions from European agriculture - An analysis of variability and drivers of emissions from field experiments," Biogeosciences, vol. 10, no. 4, pp. 2671–2682, 2013, doi: 10.5194/bg-10-2671-2013.

[5] J. Feng et al., "Impact of agronomy practices on the effects of reduced tillage systems on CH4 and N2O emissions from agricultural fields: A global meta-analysis," PLoS One, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196703.

[6] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," Jun. 01, 2021, Association for Computing Machinery. doi: 10.1145/3444690.

[7] Q. Yi et al., "Effects of nitrogen application rate, nitrogen synergist and biochar on nitrous oxide emissions from vegetable field in south China," PLoS One, vol. 12, no. 4, Apr. 2017, doi: 10.1371/journal.pone.0175325.

[8] M. E. Foltz, J. L. Zilles, and S. Koloutsou-Vakakis, "Prediction of N 2 O emissions under different field management practices and climate conditions," Science of the Total Environment, vol. 646, pp. 872–879, Jan. 2019, doi: 10.1016/j.scitotenv.2018.07.364.

[9] T. F. Schindler, S. Schlicht, and K. D. Thoben, "Towards Benchmarking for Evaluating Machine Learning Methods in Detecting Outliers in Process Datasets," Computers, vol. 12, no. 12, Dec. 2023, doi: 10.3390/computers12120253.

[10] D. Cortes, "Revisiting randomized choices in isolation forests," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.13402

[11] S. Misra, O. Osogba, and M. Powers, "Unsupervised outlier detection techniques for well logs and geophysical data," in Machine Learning for Subsurface Characterization, Elsevier, 2019, pp. 1–37. doi: 10.1016/B978-0-12-817736-5.00001-6.

[12] J. J. Michael and M. Thenmozhi, "Outlier detection in maize field using Isolation Forest: A one-class classifier," in 2023 International Conference on Networking and Communications (ICNWC), Apr. 2023, pp. 1–6. doi: 10.1109/ICNWC57852.2023.10127404.

[13] A. A. Bushra and G. Yi, "Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms," IEEE Access, vol. 9, pp. 87918–87935, 2021, doi: 10.1109/ACCESS.2021.3089036.

[14] X. Zhang, X. Wang, and Y. Chen, "Carbon Emission Prediction and Clean Industry Transformation Based on Machine Learning: A Case Study of Sichuan Province."

[15] R. A. M. San Ahmed, "Hard Voting Approach using SVM, Naïve Bays and Decision Tree for Kurdish Fake News Detection," Iraqi Journal for Computer Science and Mathematics, vol. 4, no. 3, pp. 25–33, 2023, doi: 10.52866/ijcsm.2023.02.03.003.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[17] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," International Journal of Computers and Applications, vol. 44, no. 9, pp. 875–886, 2022, doi: 10.1080/1206212X.2021.1974663.

[18] D. Navon and A. M. Bronstein, "Random Search Hyper-Parameter Tuning: Expected Improvement Estimation and the Corresponding Lower Bound," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.08170

[19] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," J. Mach. Learn. Res., vol. 13, pp. 281–305, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:15700257

[20] W. Ashiq, H. B. Vasava, U. Ghimire, P. Daggupati, and A. Biswas, "Topography controls n2o emissions differently during early and late corn growing season," Agronomy, vol. 11, no. 1, Jan. 2021, doi: 10.3390/agronomy11010187.

[21] D. Saha, B. Basso, and G. P. Robertson, "Machine learning improves predictions of agricultural nitrous oxide (N2O) emissions from intensively managed cropping systems," Environmental Research Letters, vol. 16, no. 2, Feb. 2021, doi: 10.1088/1748-9326/abd2f3.

[22] C. D. Dorich et al., "Improving N2O emission estimates with the global N2O database," Dec. 01, 2020, Elsevier B.V. doi: 10.1016/j.cosust.2020.04.006.

[23] Z. Shang et al., "Measurement of N2O emissions over the whole year is necessary for estimating reliable emission factors," Environmental Pollution, vol. 259, Apr. 2020, doi: 10.1016/j.envpol.2019.113864.

[24] A. R. Muslikh, H. A. Santoso, P. N. Andono, and A. Marjuni, "2023 International Seminar on Application for Technology of Information and Communication (iSemantic)."

[25] J. Li et al., "Feature selection: A data perspective," Dec. 01, 2017, Association for Computing Machinery. doi: 10.1145/3136625.

[26] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[27] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Irianandа, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," J. Futur. Artif. Intell. Technol., vol. 1, no. 1, pp. 23–38, 2024, doi: 10.62411/faith.2024-11.

[28] M. Safaei et al., "A systematic literature review on outlier detection inwireless sensor networks," Mar. 01, 2020, MDPI AG. doi: 10.3390/sym12030328.

[29] S. Bharti, K. K. Pattanaik, and A. Pandey, "Contextual outlier detection for wireless sensor networks," J Ambient Intell Humaniz Comput, vol. 11, no. 4, pp. 1511–1530, Apr. 2020, doi: 10.1007/s12652-019-01194-5.

[30] L. Popescu and A. S. Safta, "The Causal Relationship of Agricultural Standards, Climate Change and Greenhouse Gas Recovery," MDPI AG, Mar. 2021, p. 21. doi: 10.3390/ecas2020-08153.

[31] P. W. Beamish and V. C. Hasse, "The importance of rare events and other outliers in global strategy research," Global Strategy Journal, vol. 12, no. 4, pp. 697–713, Nov. 2022, doi: 10.1002/gsj.1437.

[32] F. M. Firnando, D. R. I. M. Setiadi, A. R. Muslikh, and S. W. Irianandа, "Analyzing InceptionV3 and InceptionResNetV2 with Data Augmentation for Rice Leaf Disease Classification," J. Futur. Artif. Intell. Technol., vol. 1, no. 1, pp. 1–11, 2024, doi: 10.62411/faith.2024-4.

[33] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier," in Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019, Association for Computing Machinery, Inc, Sep. 2019, pp. 161–168. doi: 10.1145/3338840.3355641.

[34] L. Anusha and G. S. Nagaraja, "Outlier Detection in High Dimensional Data," Int J Eng Adv Technol, vol. 10, no. 5, pp. 128–130, Jun. 2021, doi: 10.35940/ijeat.e2675.0610521.

[35] Q. Yang, J. Singh, and J. Lee, "Isolation-Based Feature Selection for Unsupervised Outlier Detection."

[36] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[37] M. Hahsler, M. Piekenbrock, and D. Doran, "Dbscan: Fast density-based clustering with R," J Stat Softw, vol. 91, 2019, doi: 10.18637/jss.v091.i01.

[38] O. Virgantara Putra, T. Harmini, and A. Saroji, "Outlier Detection On Graduation Data Of Darussalam Gontor University Using One-Class Support Vector Machine."

[39] C. Tao, T. Li, and J. Huang, "Kernel Choice in One-Class Support Vector Machines for Novelty and Outlier Detection," in Proceedings - 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 116–120. doi: 10.1109/MLBDBI51377.2020.00026.

[40] J. J. Cherian et al., "Efficient hyperparameter optimization by way of PAC-Bayes bound minimization," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.06431

[41] T. Cao Truong, "Ensemble Learning Approaches For Classification With High-Dimensional Data," Journal of Science and Technique, vol. 12, no. 01, Jun. 2023, doi: 10.56651/lqdtu.jst.v12.n1.659.ict.

[42] A. Saha, A. Chatterjee, S. Ghosh, N. Kumar, and R. Sarkar, "An ensemble approach to outlier detection using some conventional clustering algorithms," Multimed Tools Appl, vol. 80, no. 28–29, pp. 35145–35169, Nov. 2021, doi: 10.1007/s11042-020-09628-5.

[43] L. Shi and C. Zhu, "Selective Combination based on Diversity-Accuracy Balance in Outlier Ensembles," in 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Dec. 2020, pp. 1274–1281. doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00165.

[44] M. M. Singh and N. Kane, "Outlier Detection using Ensemble Learning," in 2022 6th International Conference on Information Technology (InCIT), 2022, pp. 234–239. doi: 10.1109/InCIT56086.2022.10067524.

[45] F. Mouret, M. Albughdadi, S. Duthoit, D. Kouamé, G. Rieu, and J. Y. Tourneret, "Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and sar time series," Remote Sens (Basel), vol. 13, no. 5, pp. 1–25, Mar. 2021, doi: 10.3390/rs13050956.

[46] B. Ouyang, Y. Song, Y. Li, G. Sant, and M. Bauchy, "EBOD: An ensemble-based outlier detection algorithm for noisy datasets," Knowl Based Syst, vol. 231, p. 107400, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107400.