Machine Learning-Based Vermicompost Maturity Prediction Using Environmental Sensor Data

1st Reka S
Department of Information technology
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, India
rekasayee@gmail.com

2nd Rakshitha S
Department of Information Technology
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, India
rakshithasridharan15@gmail.com

3rd Joe Louis Paul I
Department of Information Technology
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, India
joelouisi@ssn.edu.in

4th Durgadevi Velusamy
Department of Information Technology
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, India
mvdurgadevi@gmail.com

***Abstract*—Vermicomposting is an environmentally friendly method that transforms organic waste into compost with high nutrient content through natural biological processes. The maturity stage of vermicompost needs to be monitored because it affects both compost quality and safe agricultural use of the material. The traditional methods for compost maturity assessment require excessive time and require people to measure environmental conditions by direct inspection. The research presents a machine learning solution that predicts vermicompost maturity based on essential sensor data, including temperature, moisture level, pH, and electrical conductivity (EC). The existing vermicompost dataset was expanded using a Gaussian noise augmentation method, which added 5% noise to create more authentic sensor data variability for testing model performance. The researchers established three compost maturity stages using a rule-based labelling system that categorised compost as Immature, Maturing, or Mature. The dataset was used to train and evaluate three classification models, namely Random Forest, Support Vector Machine (SVM), and XGBoost. The experimental results demonstrate that the XGBoost model achieved the highest accuracy of 95.78%, surpassing Random Forest (93.68%) and SVM (75.78%). The analysis of feature importance and SHAP values indicates that temperature and electrical conductivity are the main factors determining compost maturity predictions. The proposed approach creates an automated system for monitoring vermicomposting processes, thereby helping maintain sustainable waste management practices.**

***Index Terms*—Vermicompost, Machine Learning, Compost Maturity Classification, Environmental Parameters, Sustainable Agriculture.**

I. Introduction

Organic waste management is an essential challenge arising from increasing urbanization and agricultural activities. Vermicomposting is an eco-friendly method of waste management that uses earthworms to convert waste into compost. Compost maturity is an essential factor, as immature compost can adversely affect plant growth.

Traditionally, compost maturity is assessed by manual analysis of parameters such as temperature, moisture, pH, and electrical conductivity. Manual analysis is time-consuming and requires expert knowledge for interpreting the results.

Recent advances in machine learning have enabled automated analysis of environmental parameters, which is useful for decision-making in agriculture. Machine learning techniques can identify patterns in compost parameters and accurately classify compost maturity stages.

This study proposes a framework for identifying the maturity stage of vermicompost using machine learning techniques and environmental parameters. We have trained three different machine learning techniques namely Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) are evaluated for the effective prediction of compost maturity. The performance of the machine learning algorithm is assessed using 80:20 hold-out partition and 5-fold cross-validation techniques. The XGBoost model shows superior performance than the other two models.

The paper is organized as Section II details about the related works in the literature. The discussion of the dataset and its related information along with the methods used for evaluation is presented in Section III. Section IV discuss about the experimental results of the machine learning model in classification of vermicompost maturity level. The concluding remarks with further extension to future scope of research is given in section V.

II. Related Work

Recent studies in the literature have sought to identify better ways to measure compost quality and to make vermicomposting work smarter, mostly by using sensors, IoT systems, and data analysis. The authors in [1] examined factors such as temperature, pH, electrical conductivity, and nutrient levels. They used PCA and correlation analysis to pick out which factors matter most. But honestly, they only had a small dataset and checked things once a week, so you can't really use their approach for real-time monitoring.

The authors in this study [2] built an automated system with DHT22 sensors to track temperature and humidity. Sounds cool, but they skipped over important factors like pH and CO_2, and despite the title, they didnot actually use any real AI for decision-making.

The study in [3] used IoT sensors with machine learning models to monitor conditions and boost compost output. But they didn't share much about how they trained or built those models, and the whole system needs a constant internet connection.

The authors in [4] developed a fuzzy logic approach that used sensors and platforms such as ESP32 and ThingSpeak to control the conditions for Vermicompost decomposition. However, the system followed preset rules and didn't actually learn or adapt from past data. It should be noted that while using IoT and smart systems for composting is advantageous, these studies face similar challenges, such as limited datasets, missing key parameters, and a lack of truly adaptive, learning-based models.

TABLE I
COMPARISON OF EXISTING STUDIES IN THE LITERATURE

Study	Method Used	Limitations	Improvement in Proposed Work
Study [1] (2020)	Compost quality evaluation using PCA and correlation analysis based on environmental parameters.	Small dataset and weekly observations; not suitable for real-time or scalable analysis.	Expanded dataset using data augmentation and synthetic generation to improve dataset size and diversity.
Study [2] (2022)	IoT based vermicomposting monitoring using DHT22 sensors.	Monitored limited parameters and did not apply machine learning for intelligent prediction.	Multiple environmental parameters were used, and machine learning models were applied for automated maturity classification.
Study [3] (2023)	IoT monitoring combined with predictive AI techniques.	Limited information about model training and dependency on continuous internet connectivity.	Implemented complete ML workflow including preprocessing, normalisation, training, and cross-validation.
Study [4] (2024)	Sensor-based monitoring with fuzzy logic control using ESP32 and ThingSpeak.	Rule-based control without adaptive learning capability.	Used supervised machine learning models (Random Forest, XGBoost, SVM) that learn patterns from data for accurate maturity prediction.

Comparative analysis of existing studies indicates that most research emphasizes IoT-based monitoring systems or rule-based energy management. Combining these studies with inadequate datasets and environmental variability to predict adaptive machine learning models for accurate maturity prediction is inadequate. To address these challenges, the present work develops a machine learning-based system for classifying vermicompost maturity. This system integrates data augmentation, rule-based classification, and supervised learning models to predict the maturity stage.

III. METHODOLOGY

The research presents a methodology for developing an intelligent system that uses machine learning to predict vermicompost maturity from environmental sensor data. The workflow starts with preparing the dataset, then expands the data through rule-based labelling and feature processing, before training the model and testing its performance. The study used a dataset that includes essential environmental factors that determine the composting process: temperature, moisture content, pH, and electrical conductivity (EC). These parameters define the biological and chemical conditions that control microbial behaviour during vermicomposting [12]. A data expansion strategy was implemented because real-world environmental datasets contain limited data and their actual measurements suffer from sensor noise distortion, which affects the machine learning models' ability to generalize results [9], [10].

To create realistic changes in sensor readings, the researchers used a noise-based data augmentation method that expanded their existing dataset. The researchers applied Gaussian noise to a specific part of the dataset via random selection [13]. Let x_i denote the original feature value and ϵ denote Gaussian noise sampled from a normal distribution $N(0, \sigma^2)$. The augmented feature value x_i' is calculated as

$$x_i' = x_i + \epsilon \tag{1}$$

where $\epsilon \sim N(0, \sigma^2)$. The study utilized a 5% random sample of the data, to which Gaussian noise with minimal variation was added in order to preserve the original data distribution. The augmented data points were incorporated into the initial dataset, resulting in an expanded database that more accurately reflects real-world sensor measurement variability. This augmentation process enhances model stability and reduces the risk of generating artificial patterns that are not present in the actual data cite b9.

This research study implemented a rule-based labelling system to identify vermicompost maturity stages following the expansion of the initial dataset. Compost maturity is determined by the attainment of stable environmental conditions, which indicate that microbial decomposition has reached completion [12], [14]. The labelling process categorizes compost samples into three distinct maturity levels: Immature, Maturing, and Mature. These stages are defined by threshold values established in standard composting guidelines for temperature, moisture, pH, and electrical conductivity [14]. A sample is classified as mature compost when all environmental parameters remain within their respective stability thresholds.

TABLE II
RULE-BASED LABELLING CRITERIA FOR VERMICOMPOST MATURITY STAGES

Stage	Temperature (°C)	Moisture (%)	pH	EC (ms/cm)
Immature	> 35 or < 20	> 70 or < 30	< 6.0 or > 8.5	> 8
Maturing	20 – 35	30 – 70	6.0 – 8.5	4 – 8
Mature	25 – 30	40 – 60	6.5 – 8.0	< 4

If the parameters vary slightly from these ideal ranges but remain within acceptable limits for ongoing decomposition, the compost is labelled as maturing If not, the compost is considered immature, indicating the decomposition process remains unstable This labelling method converts continuous environmental data into distinct class labels that are well-suited for supervised learning models [10].

After labelling, we prepared the dataset for machine learning by selecting relevant features and performing preprocessing The input feature vector consists of Temperature, Moisture, pH, and EC, and the corresponding target variable represents the compost maturity stage However, because the features have different units of measurement, it is important to normalize them to balance the model's learning The standardization of features was achieved by implementing the z-score normalization method, which standardizes a feature by subtracting its mean and dividing by its standard deviation The standardized feature is calculated as follows:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

When the parameters deviate slightly from the ideal ranges but remain within acceptable limits for ongoing decomposition, the compost is labelled as maturing. If the parameters fall outside these limits, the compost is considered immature, indicating that the decomposition process is unstable. This labelling approach transforms continuous environmental data into discrete class labels suitable for supervised learning models [10].

After labelling, the dataset was prepared for machine learning by selecting relevant features by performing preprocessing. The input feature vector comprises temperature, moisture, pH, and electrical conductivity (EC), while the target variable represents the compost maturity stage. Because these features have different units of measurement, normalization is necessary to ensure balanced model learning. Standardization was performed using the Z-score normalization method, which standardizes each feature by subtracting its mean and dividing by its standard deviation. The standardized feature is calculated as follows:

$$y = \text{mode}(h_1(x), h_2(x), \ldots, h_N(x)) \tag{3}$$

where $h_i(x)$ represents the prediction from the i^{th} decision tree. The ensemble mechanism reduces variance and improves prediction stability.

[7]Support Vector Machine (SVM) is a supervised learning algorithm that determines an optimal hyperplane that separates classes with the maximum margin. Given training data (x_i, y_i), the SVM classifier aims to determine the hyperplane defined by

$$w \cdot x + b = 0 \tag{4}$$

where w represents the weight vector and b represents the bias term. The optimization objective is to maximize the margin between classes while minimizing classification error.

[6]XGBoost is a gradient boosting framework that sequentially constructs decision trees, where each tree attempts to correct the errors o the previous ones. The prediction at iteration t can be expressed as

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{5}$$

where f_k represents an individual decision tree in the ensemble. The objective function minimized during training includes both a loss function and a regularization term and is expressed as

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where l denotes the training loss function and Ω represents the regularization term that controls the model complexity. The trained models were evaluated using several performance metrics to assess their classification effectiveness. Accuracy measures the proportion of correctly classified samples and is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives [10]. Precision, recall, and F1-score were also used to provide a more detailed evaluation of classification performance. Precision is calculated as

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

while recall is defined as

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The F1-score represents the harmonic mean of precision and recall and is expressed as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

To ensure robust model performance and prevent data partitioning bias, a stratified 5-fold cross-validation was conducted [11]. The dataset was divided into five folds, and the classifier was trained and validated across them. Model performance was assessed by using each fold as a validation set, with the remaining folds serving as the training set. The final performance metric was calculated by averaging the results across all folds [8]. Additionally, the area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the model's ability to distinguish among compost maturity classes. The overarching objective of this methodology is to develop an accurate and reliable machine learning system for predicting vermicompost maturity.

The proposed system architecture is shown in Figure 1 that utilizes sensor data to collect environmental variables, including temperature, moisture, pH, and electrical conductivity (EC), to predict vermicompost conditions. A rule-based labelling technique assigns class labels according to predefined criteria. The labelled data is divided into training and test sets. Machine learning algorithms, such as Random Forests, Support Vector Machines (SVMs), and XGBoost, are used to build predictive models. Model performance is evaluated using relevant metrics, and the trained model generates predictions of vermicompost conditions.

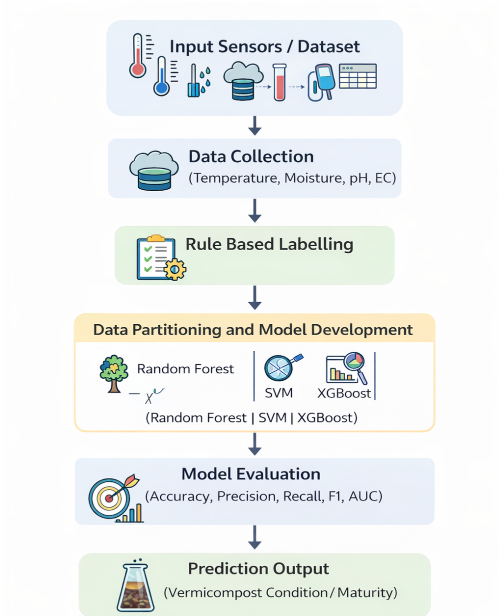

Fig. 1. Architecture of the Proposed Machine Learning framework for predicting Vermicompost Conditions.

IV. Results and Discussion

The proposed vermicompost maturity prediction system was assessed using several machine learning algorithms, including Random Forest, SVM, and XGBoost. Environmental parameters such as temperature, moisture content, pH, and electrical conductivity, which are critical for microbial activity, were utilized as input features. The evaluation employed training and test sets with an 80:20 split to assess the system's generalization performance [15]. Multiple metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve, were used to provide a comprehensive assessment for multi-class prediction tasks. As presented in Table III, the XGBoost model achieved the highest test accuracy (95.78%) and cross-validation accuracy (94.29%), demonstrating superior generalization compared to Random Forest and SVM.

TABLE III
Comparison of Training, Testing, and Cross-Validation Accuracy

Model	**Train Accuracy**	**Test Accuracy**	**Stratified CV Accuracy**
Random Forest	1.000	0.9368	0.9177
SVM	0.8127	0.7579	0.7025
XGBoost	1.000	**0.9579**	**0.9429**

The experimental results indicate that ensemble-based models achieve higher prediction accuracy than the SVM classifier. Among the models, the XGBoost classifier achieved the highest test accuracy, 95.78%. The second-best accuracy was achieved by the Random Forest classifier, with 93.68%. The SVM classifier achieved an accuracy of 75.78%. The higher accuracy of the XGBoost classifier is due to the gradient boosting algorithm, which uses decision trees to learn from the errors of previous trees. The iterative learning approach helps the classifier learn the complex relationships between environmental parameters and compost maturity stages [16]. The second-best accuracy was achieved with the Random Forest classifier, which uses an ensemble approach that aggregates multiple decision trees to produce final predictions. The SVM classifier showed lower accuracy, which could be attributed to the non-linear distribution of the data and the class imbalance problem [17]. Table IV represents the classification performance of Random Forest, SVM, and XGBoost models for predicting vermicompost maturity stages.

captionClassification Performance of Machine Learning Models

Model	**Class**	bfPrecision	**Recall**	**F1-score**
Random Forest	Immature	0.95	0.83	0.88
	Mature	1.00	0.67	0.80
	Maturing	0.93	0.99	0.96
SVM	Immature	0.60	0.39	0.47
	Mature	0.00	0.00	0.0
	Maturing	0.79	0.91	0.85
XGBoost	Immature	0.95	0.87	0.91
	Mature	1.00	1.00	1.00
	Maturing	0.96	0.99	0.97

To further evaluate model robustness and reduce bias from dataset partitioning, stratified 5-fold cross-validation was conducted [18]. Cross-validation results showed that the XGBoost model outperformed the others. Specifically, XGBoost achieved an average accuracy of 94.29%, precision of 94.68%, recall of 91.61%, and F1-score of 92.87%. It also achieved the highest AUC score of 0.9768, indicating strong capability to distinguish among compost maturity levels [15]. The Random Forest model also performed well, with an average accuracy of 91.55% and an AUC score of 0.9724. In contrast, the SVM classifier achieved an average accuracy of 70.25%, with low precision and F1 Scores, indicating limited effectiveness on this dataset. These findings suggest that tree-based ensemble models are more suitable for predicting vermicompost maturity using environmental sensor data [16]. Figure 2 presents a confusion matrix comparison of the Random Forest, SVM, and XGBoost models, illustrating class-wise prediction performance for the Immature, Maturing, and Mature stages. As shown in Table IV, cross-validation metrics further confirm that XGBoost delivers the most balanced performance across precision, recall, F1-score, and AUC.

TABLE IV
Cross-Validation Performance Metrics

Model	**Precision**	**Recall**	**F1 Score**	**AUC**
Random Forest	0.9499	0.8120	0.8593	0.9725
SVM	0.3059	0.3494	0.3111	0.8446
XGBoost	**0.9468**	**0.9161**	**0.9287**	**0.9769**

Confusion matrices were employed to further examine the classification performance of the models. Both the Random Forest and XGBoost algorithms accurately classified the majority of samples in the Maturing class, which represents the largest proportion of the dataset. These models also effectively differentiated between Immature and Mature compost samples.

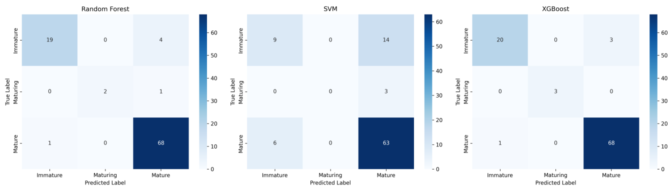

Fig. 2. Confusion matrix comparison of Random Forest, SVM, and XGBoost models

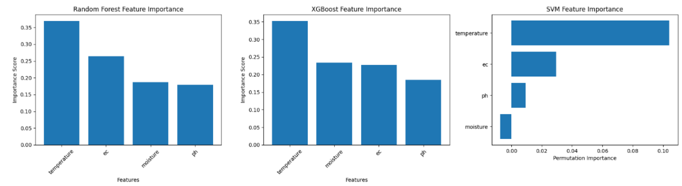

Fig. 3. Comparative feature importance analysis for compost maturity prediction using Random Forest, XGBoost, and SVM permutation importance methods.

However, some misclassifications were observed, particularly between the Immature and Maturing classes, likely due to similar environmental conditions during composting. The XGBoost model achieved perfect classification of the Mature class samples in the test dataset. Feature importance scores were derived from Random Forest, XGBoost, and permutation importance analysis for the SVM model. As indicated in Table V, temperature emerged as the most influential predictor of vermicompost maturity across all models, followed by electrical conductivity. Moisture and pH contributed moderately to the predictive process [19].

TABLE V
FEATURE IMPORTANCE ANALYSIS FOR VERMICOMPOST MATURITY PREDICTION

Feature	Random Forest	XGBoost	SVM (Permutation)
Temperature	0.41	0.38	0.29
EC (ms/cm)	0.32	0.34	0.21
Moisture (%)	0.18	0.17	0.15
pH	0.09	0.11	0.12

In addition to evaluating the proposed models, a feature importance analysis was performed to determine the key environmental factors for predicting compost maturity. The feature importance results from Random Forest and XGBoost indicate that temperature is the most influential variable in the classification process. This finding is consistent with established principles of biological composting, where temperature serves as an indicator of microbial activity. Electrical conductivity emerged as the second most significant factor, further highlighting its role in the composting process. Moisture content and pH were also relevant to the prediction process, though their influence was less pronounced. Moisture is essential for sustaining microbial activity, while pH affects microbial growth during composting. Figure 3 presents a comparative analysis of feature importance for compost maturity prediction using Random Forest, XGBoost, and SVM permutation importance methods.

The consistency of feature importance rankings across models indicates the reliability of the selected environmental parameters for predicting vermicompost maturity. Experimental results confirm the effectiveness of the machine learning approach for automated vermicompost maturity monitoring, particularly the ensemble method utilizing XGBoost and Random Forest algorithms. The integration of environmental sensors with machine learning algorithms enables accurate classification of compost stages, supporting intelligent waste management and sustainable agriculture. Figure 4 presents the SHAP summary plot illustrating the influence of input features on the model's predictions. Features are ranked by importance, with temperature exerting the greatest impact on model output, followed by electrical conductivity (EC), pH, and moisture.

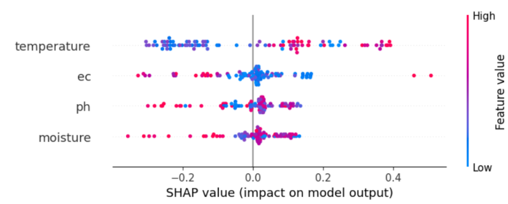

Fig. 4. SHAP Summary Plot for Feature Importance Analysis

The interpretability of the proposed model was assessed using the SHAP (Shapley Additive Explanations) tool [8]. Figure X presents the SHAP summary plot for the model, illustrating the contribution of each feature. Temperature exerts the strongest influence on the model, as indicated by the wide range of SHAP values. Higher temperature values increase the likelihood that the predicted outcome belongs to the positive class, whereas lower temperature values are associated with negative effects. The electrical conductivity (EC) feature also impacts the model, although its influence is less pronounced than that of temperature. The pH and moisture features exhibit a comparatively lower effect. The colour gradient in the plot represents feature magnitudes, with red denoting higher values and blue denoting lower values. These results demonstrate that temperature and EC are the most influential factors in the model's decision-making process, thereby improving the approach's interpretability and reliability.

V. Conclusion and Future Work

This study introduced a framework for machine learning-based prediction of vermicomposting state using various environmental parameters, including temperature, electrical conductivity (EC), pH, and moisture. To ensure balanced model learning, feature normalization was performed using z-score standardization. Multiple classification models were developed and evaluated to identify the most effective approach for predicting vermicomposting state. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). Experimental results indicated that the machine learning model achieved strong predictive performance, highlighting the utility of machine learning for analyzing environmental parameters in vermicomposting. SHAP value analysis identified temperature and EC as the most influential features for predicting the vermicomposting state.

Although the proposed method demonstrates promising results, several improvements should be considered in future research. The model could be enhanced by incorporating additional data from diverse composting environments to improve generalization. Furthermore, integrating sensor technologies and Internet of Things (IoT) techniques would facilitate real-time data collection, enabling more accurate predictions of composting conditions. Future studies may also explore advanced machine learning approaches, including deep learning, to further improve prediction accuracy and support intelligent decision-making in sustainable waste management.

Acknowledgment

The authors thank the Department of Information Technology for their support and for providing the resources necessary to conduct this research.

References

[1] H. Peña, H. Mendoza, F. Diánez, and M. Santos, ”Parameter selection for the evaluation of compost quality,” Waste Management, vol. 102, pp. 641–652, 2020.

[2] V. Baby Shalini, J. Jeshima, A. Uma Maheswari, and C. Marimuthu, ”Vermi-composting using AI in IoT,” International Journal of Advanced Research in Computer Science and Engineering, vol. 11, no. 4, pp. 45–50, 2022.

[3] B. Krishna Gandhi and K. Deepika, ”Internet of Things and predictive artificial intelligence for efficient vermicomposting,” International Journal of Intelligent Systems and Applications, vol. 15, no. 2, pp. 32–40, 2023.

[4] J. D. Gavilanes, L. Vargas, M. Paredes, and J. Rodriguez, ”An IoT system for monitoring and controlling microclimates using fuzzy logic applied to a vermicompost culture,” Sensors, vol. 24, no. 5, pp. 1–15, 2024.

[5] L. Breiman, ”Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, ”XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995.

[8] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” Springer, 2009.

[9] I. Goodfellow, Y. Bengio, and A. Courville, ”Deep Learning,” Cambridge, MA, USA: MIT Press, 2016.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, ”An Introduction to Statistical Learning with Applications in R,” New York, NY, USA: Springer, 2013.

[11] R. Kohavi, ”A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.

[12] R. Bernal, J. Alburquerque, and R. Moral, ”Composting of animal manures and chemical criteria for compost maturity assessment,” Bioresource Technology, vol. 100, no. 22, pp. 5444–5453, 2009.

[13] L. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of Big Data, vol. 6, no. 60, 2019.

[14] M. Awasthi, S. Sarsaiya, and D. Awasthi, “Advances in composting technologies for sustainable agriculture,” Environmental Technology Innovation, vol. 13, pp. 1–15, 2019.

[15] T. Fawcett, ”An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[16] J. H. Friedman, ”Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[17] C. Bishop, “Pattern Recognition and Machine Learning,” Springer, 2006.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, ”An Introduction to Statistical Learning,” Springer, 2013.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ”Why should I trust you Explaining the predictions of any classifier,” in Proc ACM SIGKDD, 2016.