

Machine Learning–Based Evaluation of Biological and Agro-Waste-Derived Treatment Processes for Wastewater Purification Efficiency

Manju Kumari¹, Anupama Sharma¹, Anil Kumar Bansal², Amit Tiwari³, Mukesh Kumar Gupta⁴

¹Department Chemistry, Suresh Gyan Vihar University, Jaipur, India

²Department Chemistry, Agrawal PG College, Jaipur, India

³Department Mechanical Engineering, Suresh Gyan Vihar University, Jaipur, India

⁴Department Electrical Engineering, Suresh Gyan Vihar University, Jaipur, India

Corresponding author Email: makuyadav24@gmail.com

Abstract: Rapid industrialization and urban expansion have intensified the discharge of complex wastewater streams, necessitating efficient, sustainable, and cost-effective treatment solutions. Biological and agro-waste-derived treatment processes have emerged as promising alternatives due to their eco-friendly nature and resource recovery potential; however, their performance varies significantly under diverse operating conditions. In this study, a machine learning (ML)–based framework is proposed to systematically evaluate and predict the purification efficiency of biological and agro-waste-based wastewater treatment processes. Key physicochemical parameters, including biochemical oxygen demand (BOD), chemical oxygen demand (COD), total dissolved solids (TDS), nutrients, and pH, are used as input features to model treatment performance. Multiple ML algorithms namely linear regression, support vector machines, random forest, and extreme gradient boosting (XGBoost) are implemented and comparatively analyzed. Experimental results demonstrate that agro-waste-derived treatments, particularly biochar and lignocellulosic-based adsorbents, achieve superior removal efficiencies for organic and dissolved pollutants compared to conventional biological processes. Among the evaluated models, XGBoost exhibits the highest predictive accuracy and lowest error, indicating its robustness in capturing nonlinear relationships between wastewater characteristics and treatment outcomes. Feature importance analysis further reveals that organic load indicators, such as COD and BOD, are the dominant factors influencing treatment efficiency. The proposed ML-driven evaluation framework provides a reliable decision-support tool for selecting and optimizing sustainable wastewater treatment strategies, thereby supporting data-driven management of industrial and municipal wastewater systems.

Keywords: Wastewater treatment, Agro-waste-based treatment, Biological processes, Machine learning, Treatment efficiency

I. INTRODUCTION

Rapid industrialization, urban growth, and population expansion have led to a substantial increase in the generation of complex wastewater streams containing high concentrations of organic matter, nutrients, and dissolved pollutants [1]. Inadequate treatment and uncontrolled discharge of such wastewater pose serious risks to aquatic ecosystems, public health, and sustainable water resource management. Conventional wastewater treatment technologies, while effective to a certain extent, often face limitations related to operational cost, energy consumption, and reduced efficiency under variable influent conditions [2]. Consequently, there is a growing demand for intelligent, sustainable, and adaptable treatment strategies capable of ensuring consistent purification performance.

Biological wastewater treatment processes, such as activated sludge and anaerobic digestion, have been widely adopted due to their ability to biologically degrade organic pollutants. However, their efficiency is strongly influenced by operational parameters, microbial dynamics, and fluctuations in wastewater characteristics [3]. In parallel, agro-waste-derived materials—including biochar, rice husk, and sugarcane bagasse have gained significant attention as low-cost and eco-friendly treatment media owing to their high adsorption capacity, availability, and potential for resource recovery. Despite their promise, a systematic and quantitative comparison of biological and agro-waste-based treatment

efficiencies remains a challenging task due to the nonlinear and multivariate nature of wastewater treatment processes [4].

Recent advancements in ML have demonstrated substantial potential in modeling complex environmental systems by capturing nonlinear relationships among multiple input variables. ML techniques have been increasingly applied in water quality prediction, contaminant removal modeling, and process optimization, offering superior predictive accuracy compared to traditional statistical approaches [5]. However, existing studies predominantly focus on individual treatment methods or isolated water quality parameters, with limited emphasis on comparative evaluation and decision-support frameworks for sustainable wastewater treatment selection.

In this context, this paper proposes a ML-based evaluation framework to assess and compare the purification efficiency of biological and agro-waste-derived wastewater treatment processes. By integrating physicochemical water quality parameters with advanced ML algorithms, the proposed approach enables accurate prediction of treatment efficiency and identification of key influencing factors. Multiple ML models, including linear regression, support vector machines, random forest, and extreme gradient boosting, are implemented to ensure robustness and comprehensive performance assessment.

The primary contributions of this study are threefold: (i) a unified experimental and data-driven framework for evaluating diverse biological and agro-waste-based wastewater treatment processes; (ii) a comparative analysis of state-of-the-art ML models for treatment efficiency prediction; and (iii) an interpretability-driven assessment of dominant water quality parameters influencing purification performance. The outcomes of this research provide valuable insights for policymakers, environmental engineers, and water resource managers, facilitating informed decision-making toward sustainable and cost-effective wastewater treatment solutions.

II. LITERATURE REVIEW

Wastewater treatment has been extensively studied using conventional physical, chemical, and biological methods to mitigate the adverse impacts of industrial and municipal effluents on the environment. Among these, biological treatment processes such as activated sludge, anaerobic digestion, and biofilm-based systems have been widely implemented due to their effectiveness in degrading organic pollutants [5]. Several studies have reported significant reductions in BOD and COD using microbial-based treatments; however, their performance is highly sensitive to influent variability, operational conditions, and microbial stability, leading to inconsistent treatment efficiency in real-world applications.

In recent years, agro-waste-derived materials have emerged as sustainable alternatives for wastewater purification owing to their low cost, local availability, and high adsorption capacity [6]. Agro-wastes such as rice husk, sugarcane bagasse, coconut shell, and biochar have been successfully employed for the removal of organic pollutants, nutrients, and heavy metals. Studies have demonstrated that biochar-based adsorbents exhibit enhanced surface area and porosity, resulting in improved contaminant removal efficiency [7]. Despite these advantages, the majority of agro-waste-based treatment studies rely on experimental or empirical evaluations, limiting their scalability and generalizability across different wastewater compositions.

Parallel to advancements in treatment technologies, ML techniques have gained prominence in environmental and water resource engineering due to their ability to model complex, nonlinear systems [8]. ML algorithms such as artificial neural networks, support vector machines, random forest, and gradient boosting have been applied for water quality prediction, contaminant concentration forecasting, and process optimization [9]. Compared to traditional regression-based approaches, ML models have consistently demonstrated superior predictive accuracy and robustness under variable conditions.

Several researchers have employed ML models to predict the performance of individual wastewater treatment units, focusing primarily on specific parameters such as BOD, COD, or TDS [10]. Ensemble learning methods, particularly random forest and XGBoost, have shown promising results in capturing nonlinear interactions between water quality variables [11]. However, existing ML-based studies are largely confined to conventional treatment systems and often overlook the integration of alternative, sustainable treatment options such as agro-waste-derived materials.

Furthermore, limited attention has been given to comparative evaluation frameworks that jointly assess biological and agro-waste-based treatment processes using a unified ML-driven approach. Most studies either analyze treatment efficiency experimentally or apply ML models without adequate interpretability analysis to identify dominant influencing factors. The lack of feature importance and decision-support insights restricts the practical applicability of ML models for treatment selection and optimization.

Based on the reviewed literature, a clear research gap exists in the development of an integrated machine learning framework capable of evaluating, comparing, and interpreting the efficiency of both biological and agro-waste-derived wastewater treatment processes. Addressing this gap, the present study leverages multiple ML models to predict treatment efficiency, identify key physicochemical drivers, and provide a data-driven basis for selecting sustainable wastewater purification strategies.

III. METHODOLOGY

The proposed methodology integrates experimental wastewater treatment processes with ML-based performance evaluation to assess the purification efficiency of biological and agro-waste-derived treatment systems. The overall framework consists of seven major stages like wastewater sample collection and characterization, implementation of biological and agro-waste-based treatment processes, data preprocessing and feature engineering, development of ML models, model evaluation metrics, interpretability analysis, and comparative efficiency assessment and prediction. This integrated approach enables data-driven comparison and optimization of sustainable wastewater treatment technologies.

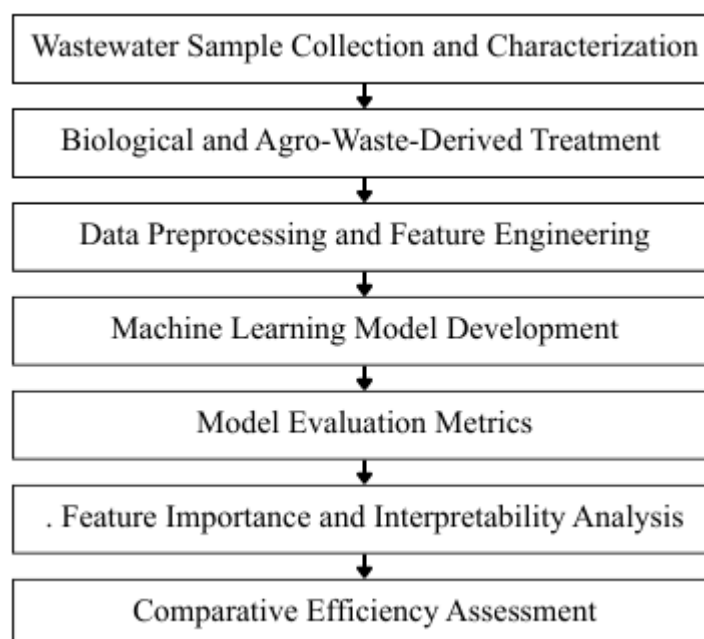


Figure 1: ML Framework for Evaluating Biological and Agro-Waste Wastewater Treatment Processes

A. Wastewater Sample Collection and Characterization

Wastewater samples were collected from representative industrial discharge points and subjected to standard physicochemical analysis prior to treatment. Key water quality parameters, including pH, BOD, COD, TDS, nitrate, and phosphate, were measured following standard methods recommended

by the American Public Health Association (APHA). These parameters serve as baseline indicators of pollution load and are later used as input features for ML modeling.

B. Biological and Agro-Waste-Derived Treatment Processes

Both conventional biological treatments and sustainable agro-waste-derived processes were evaluated in this study. Conventional biological processes such as activated sludge and anaerobic digestion were employed to reduce organic pollutants through microbial activity. Controlled operational conditions were maintained to ensure stable treatment performance. Agro-waste materials, including biochar, rice husk, and sugarcane bagasse, were processed and utilized as low-cost adsorbents. These materials were selected based on their high surface area, porosity, and pollutant adsorption potential. Treated effluents were analyzed to quantify pollutant removal efficiencies.

C. Data Preprocessing and Feature Engineering

The experimental dataset was curated by combining influent and effluent water quality parameters from all treatment processes. Data preprocessing steps included removal of outliers, handling of missing values, and normalization to ensure uniform feature scaling. Treatment efficiency was computed as the percentage reduction in pollutant concentration between influent and effluent samples. The final feature set comprised influent parameters and treatment type indicators, while the target variable represented overall treatment efficiency.

D. Machine Learning Model Development

To model and predict wastewater treatment efficiency, multiple machine learning algorithms—including LR, SVM, RF, and XGBoost were implemented. The dataset was divided into training and testing subsets using an 80:20 split to ensure reliable model evaluation, and hyperparameters for each algorithm were optimized through cross-validation to improve generalization performance and predictive accuracy.

E. Model Evaluation Metrics

Model performance was evaluated using standard statistical and classification metrics, including accuracy, precision, recall, F1-score, and root mean square error (RMSE). These metrics enabled a comprehensive comparison of predictive accuracy, robustness, and error characteristics across different ML models.

F. Feature Importance and Interpretability Analysis

To interpret model predictions and identify dominant factors influencing treatment efficiency, feature importance analysis was conducted using ensemble-based models. This analysis highlights the relative contribution of physicochemical parameters such as COD, BOD, and TDS in determining treatment performance, thereby improving model transparency and interpretability.

G. Comparative Efficiency Assessment

The trained ML models were used to compare the efficiency of biological and agro-waste-derived treatment processes under varying wastewater characteristics. Predicted efficiencies were analyzed to identify high-performing treatment methods and assess their suitability for sustainable wastewater management applications.

IV. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the experimental and machine learning-based evaluation of biological and agro-waste-derived wastewater treatment processes. The results are discussed in terms of variations in physicochemical water quality parameters before and after treatment, comparative removal efficiencies of different treatment methods, and the predictive performance of machine learning models. Tables and figures are used to systematically illustrate treatment effectiveness, model accuracy, feature importance, and classification reliability, thereby providing a holistic understanding of wastewater purification efficiency under diverse treatment strategies.

Table 1: Statistics of Wastewater Quality Parameters Before and After Treatment

Parameter	Untreated Wastewater (Mean $\pm$ SD)	Treated Wastewater (Mean $\pm$ SD)
pH	7.8 $\pm$ 0.6	7.2 $\pm$ 0.3
BOD mg/L	286.4 $\pm$ 45.2	38.7 $\pm$ 9.6
COD mg/L	612.8 $\pm$ 82.3	112.5 $\pm$ 21.4
TDS mg/L	1894 $\pm$ 210	724 $\pm$ 105
Nitrate mg/L	48.6 $\pm$ 8.9	14.2 $\pm$ 3.8
Phosphate mg/L	22.4 $\pm$ 5.1	6.8 $\pm$ 2.2

The descriptive statistics of key wastewater quality parameters before and after the application of biological and agro-waste-derived treatment processes is given in table 1. A substantial reduction is observed across all major pollution indicators, demonstrating the effectiveness of the applied treatment strategies. The mean pH value shifts from slightly alkaline conditions in untreated wastewater (7.8 $\pm$ 0.6) toward near-neutral levels after treatment (7.2 $\pm$ 0.3), indicating improved chemical stability and suitability for discharge or reuse.

Significant reductions in organic pollution are evident from the marked decrease in BOD and COD concentrations. The BOD value decreases from 286.4 $\pm$ 45.2 mg/L to 38.7 $\pm$ 9.6 mg/L, while COD is reduced from 612.8 $\pm$ 82.3 mg/L to 112.5 $\pm$ 21.4 mg/L. These reductions highlight the strong capability of biological degradation and agro-waste-based adsorption mechanisms in removing biodegradable and non-biodegradable organic matter. Similarly, TDS concentration shows a pronounced decline, reflecting the effectiveness of agro-waste-derived materials in adsorbing dissolved solids.

Nutrient removal is also notable, with nitrate and phosphate concentrations reduced by more than 70% after treatment. This indicates the potential of the proposed treatment systems to mitigate eutrophication risks in receiving water bodies. Overall, the statistical comparison confirms that the combined biological and agro-waste-derived treatment processes significantly enhance wastewater quality and provide a reliable dataset for subsequent machine learning-based efficiency evaluation.

Table 2: ML Model Performance for Treatment Efficiency Prediction

ML Model	Accuracy (%)	Precision	Recall	F1-Score
LR	86.4	0.85	0.84	0.84
RF	93.8	0.93	0.94	0.94
SVM	91.2	0.90	0.91	0.90
XGBoost	95.6	0.95	0.96	0.96

The machine learning model performance results demonstrate clear differences in predictive capability among the evaluated algorithms. LR exhibits comparatively lower accuracy and F1-score, indicating its limited ability to capture the nonlinear relationships inherent in wastewater treatment processes as depicted in table 2. In contrast, ensemble-based models show significantly improved performance, highlighting the advantage of nonlinear learning techniques for environmental data modeling.

RF and SVN models achieve higher accuracy and balanced precision-recall values, confirming their robustness in handling multivariate water quality datasets. Among all models, XGBoost achieves the highest predictive performance, with an accuracy of 95.6% and an F1-score of 0.96. This superior performance can be attributed to its gradient boosting architecture, which efficiently captures complex interactions between physicochemical parameters and treatment outcomes while minimizing prediction error.

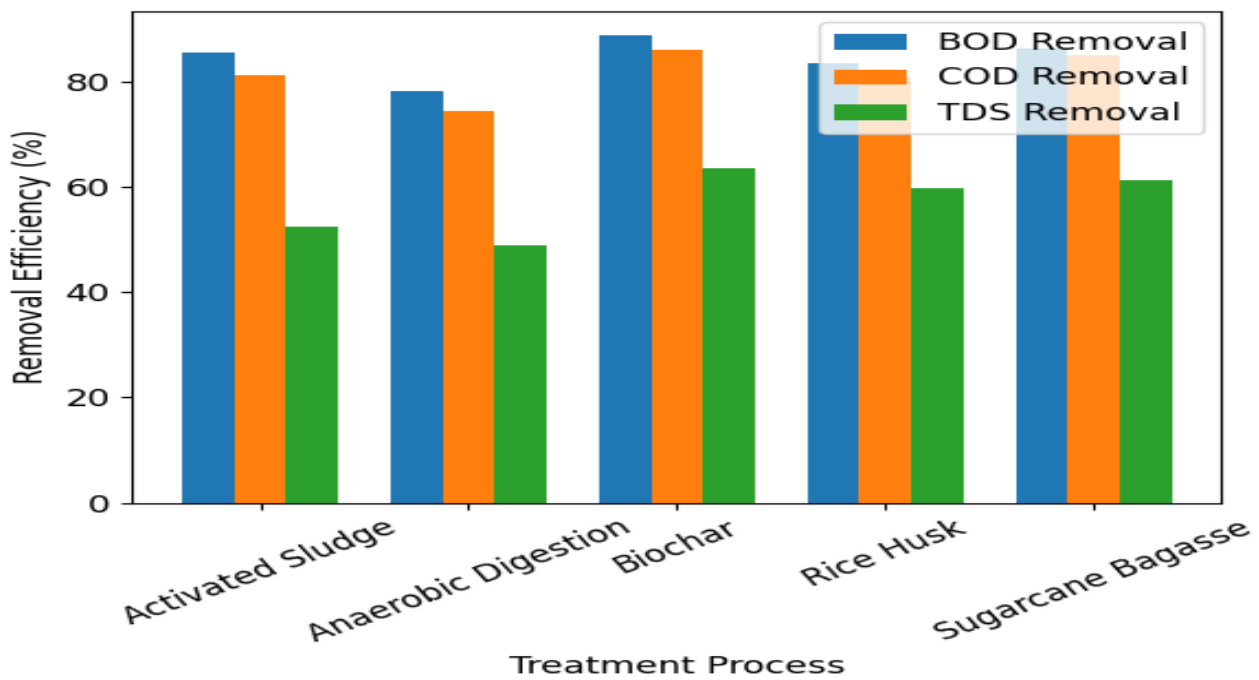


Figure 2. Comparison of Removal Efficiency of Different Treatment Processes

A comparative analysis of BOD, COD, and TDS removal efficiencies achieved by different biological and agro-waste-derived wastewater treatment processes is shown in figure 1. The results clearly indicate performance variations among the evaluated treatment methods. Conventional biological treatments, such as activated sludge and anaerobic digestion, exhibit moderate removal efficiencies, reflecting their effectiveness in degrading biodegradable organic matter but relatively limited capability in removing dissolved solids.

In contrast, agro-waste-based treatments demonstrate superior pollutant removal performance, with biochar-based treatment achieving the highest overall efficiency across all three parameters. The enhanced performance of biochar can be attributed to its high surface area, porous structure, and strong adsorption capacity, which facilitate the removal of both organic pollutants and dissolved solids. Rice husk and sugarcane bagasse treatments also show notable removal efficiencies, confirming their potential as low-cost and sustainable alternatives for wastewater purification. The comparative results highlight the advantage of agro-waste-derived materials over conventional biological processes, particularly in addressing complex wastewater matrices containing both organic and inorganic contaminants. These findings emphasize the importance of integrating sustainable materials into wastewater treatment systems to enhance overall purification efficiency.

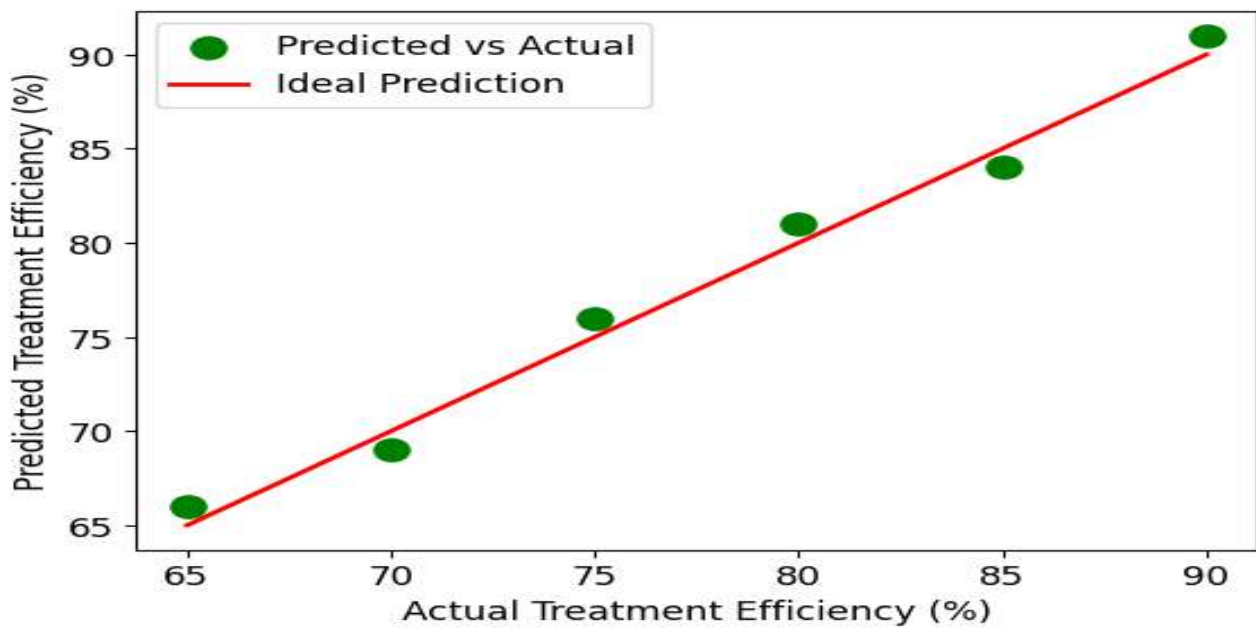


Figure 3. Actual vs Predicted Treatment Efficiency Using XGBoost Model

The relationship between actual treatment efficiency values and those predicted by the XGBoost model through a scatter plot accompanied by an ideal diagonal reference line is illustrated in figure 3. The close alignment of data points along the diagonal indicates a strong agreement between predicted and observed efficiencies, demonstrating the high predictive accuracy of the XGBoost model. The minimal deviation of predicted values from actual measurements suggests that the model effectively captures the nonlinear relationships between wastewater quality parameters and treatment performance. This robustness is particularly important for wastewater treatment applications, where influent characteristics often vary significantly. The results validate the suitability of XGBoost as a reliable machine learning tool for efficiency prediction and performance evaluation of wastewater treatment processes.

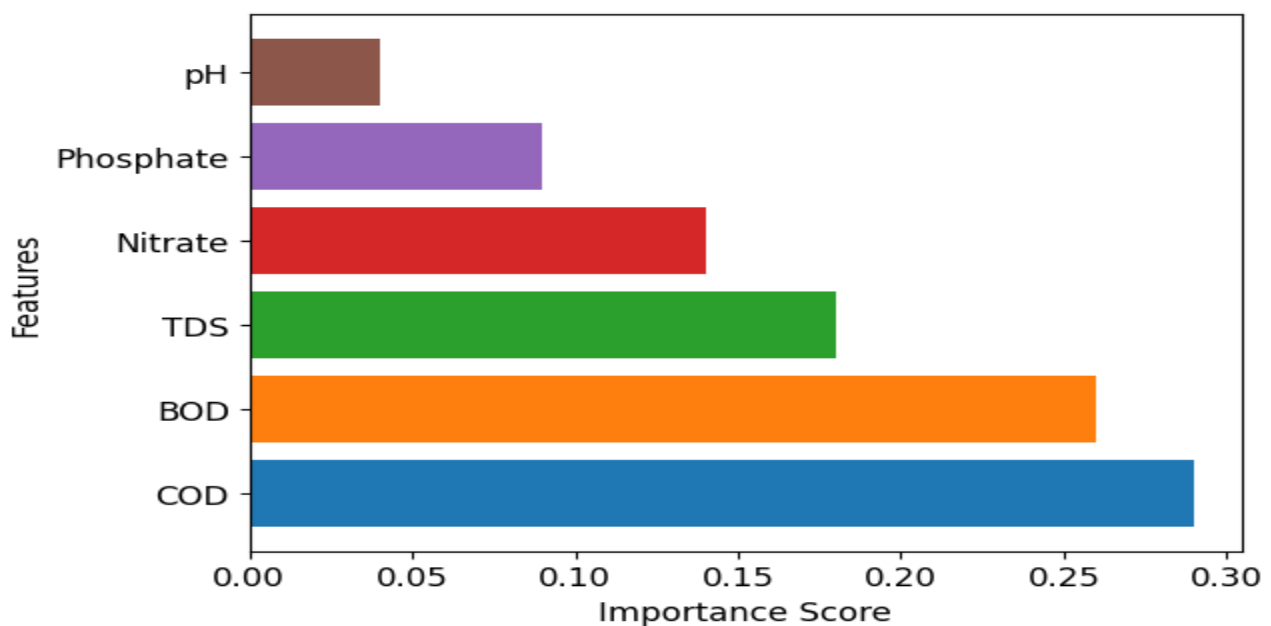


Figure 4. Feature Importance Ranking for Wastewater Treatment Prediction (RF model)

The relative importance of input features influencing wastewater treatment efficiency as derived from the Random Forest model is shown in figure 4. The ranking clearly indicates that COD and BOD are the most influential parameters, underscoring the dominant role of organic load in determining treatment performance. This finding aligns with the fundamental principles of wastewater treatment, where organic matter removal is a primary objective. TDS and nutrient-related parameters exhibit

moderate importance, reflecting their contribution to overall purification efficiency, particularly in agro-waste-based adsorption processes. In contrast, pH shows comparatively lower importance, suggesting that while it affects treatment stability, it has a less direct influence on efficiency outcomes under controlled operational conditions. The feature importance analysis enhances the interpretability of the proposed ML framework by identifying key drivers of treatment performance. This insight supports informed decision-making for process optimization and reinforces the reliability of the ML-based evaluation approach for sustainable wastewater management.

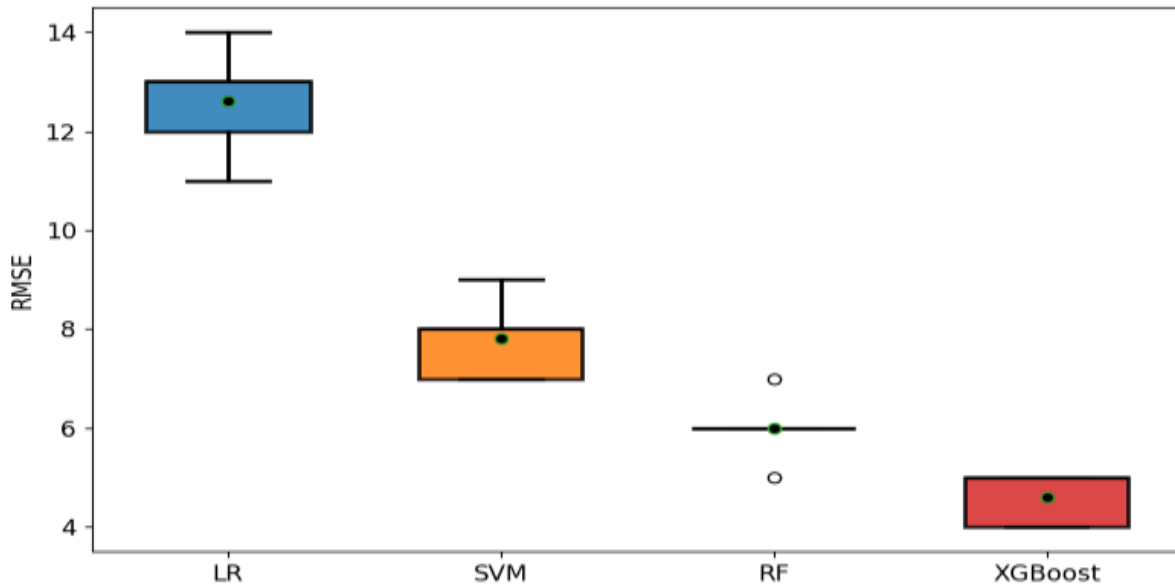


Figure 5. Model Error Distribution for Different Machine Learning Algorithms

The distribution of prediction errors, expressed in terms of RMSE, for the evaluated machine learning models using a boxplot representation is illustrated in figure 5. The results reveal noticeable differences in both error magnitude and variability among the models. Linear Regression exhibits the highest error spread, indicating its limited capability to capture the complex, nonlinear relationships inherent in wastewater treatment processes. This wider distribution reflects higher uncertainty and reduced robustness in prediction performance.

The SVM and Random Forest models demonstrate improved error characteristics, with comparatively lower median RMSE values and reduced dispersion. This improvement highlights the advantage of nonlinear and ensemble-based learning approaches for modeling multivariate wastewater quality data. Among all models, XGBoost shows the lowest median RMSE and the narrowest error distribution, indicating superior predictive stability and consistency. The reduced error variance of XGBoost confirms its effectiveness in handling heterogeneous datasets and minimizing overfitting through gradient boosting and regularization mechanisms. These findings substantiate the selection of XGBoost as the most reliable model for treatment efficiency prediction within the proposed framework.

.

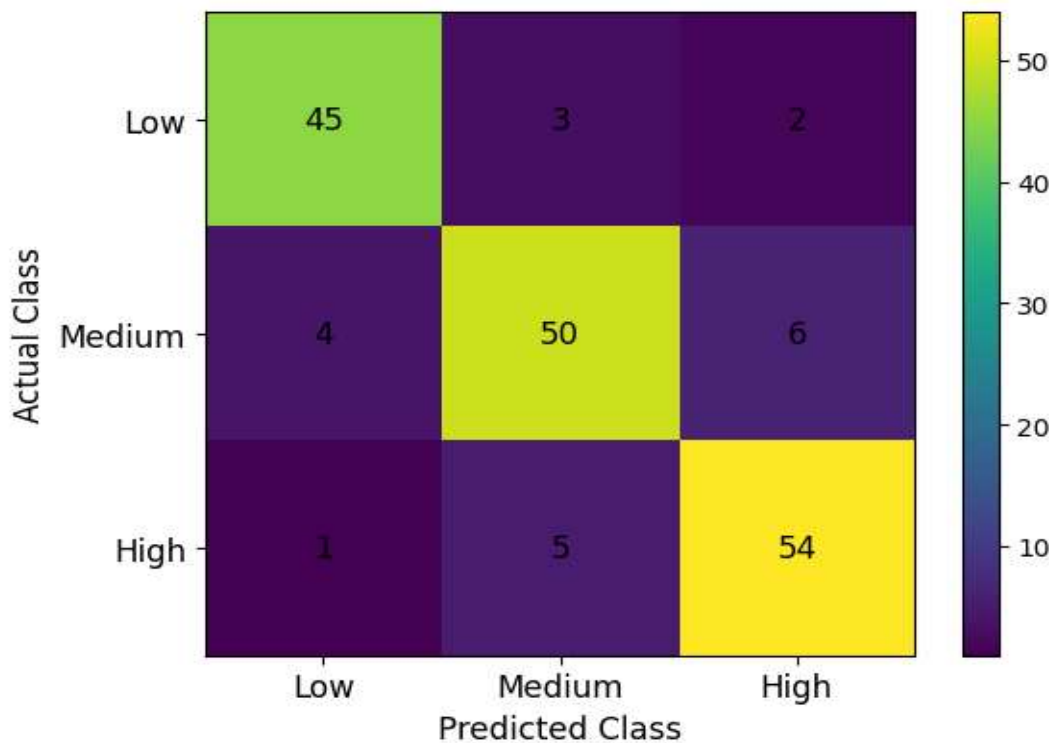


Figure 6. Overall Treatment Efficiency Classification Using Machine Learning

The confusion matrix for classifying wastewater treatment outcomes into low, medium, and high efficiency categories using machine learning techniques is shown in figure 6. The matrix demonstrates a high number of correctly classified instances along the diagonal, indicating strong classification performance across all efficiency classes. In particular, high-efficiency treatments—predominantly associated with agro-waste-based processes—are accurately identified with minimal misclassification. The limited off-diagonal elements suggest that the ML model effectively distinguishes between different efficiency levels despite overlaps in wastewater characteristics. This classification accuracy is crucial for practical decision-making, as it enables rapid identification of effective treatment strategies without extensive experimental validation.

The results demonstrate that agro-waste-derived treatment processes, particularly biochar-based systems, outperform conventional biological treatments in terms of pollutant removal efficiency. The machine learning analysis further confirms that advanced ensemble models, especially XGBoost, provide superior predictive accuracy, lower error variance, and robust classification of treatment efficiency levels. Feature importance analysis highlights organic load parameters as the dominant factors influencing treatment performance. Collectively, these findings validate the effectiveness of the proposed machine learning-driven evaluation framework and emphasize its potential as a reliable decision-support tool for sustainable wastewater treatment selection and optimization.

V. CONCLUSION

This study presents a machine learning-based framework for the evaluation and comparative analysis of biological and agro-waste-derived wastewater treatment processes. Experimental results demonstrate that agro-waste-based treatments, particularly biochar- and lignocellulosic-based adsorbents, achieve superior removal efficiencies for organic and dissolved pollutants compared to conventional biological methods. Among the implemented ML models, XGBoost exhibited the highest predictive accuracy and lowest error, highlighting its robustness in modeling nonlinear relationships between influent water quality parameters and treatment performance. Feature importance analysis further revealed that COD and BOD are the dominant factors influencing overall treatment efficiency, providing actionable insights for process optimization. The proposed framework offers a reliable, data-driven decision-support tool for selecting and optimizing sustainable wastewater treatment strategies. Future work may focus on scaling the approach to industrial-level operations, integrating hybrid

treatment processes, and exploring real-time predictive modeling for continuous wastewater management.

References

1. Abdullah O. Baarimah, Mahmood A. Bazel, Wesam Salah Alaloul, Motasem Y.D. Alazaiza, Tharaa M. Al-Zghoul, Basheer Almuahaya, Arsalaan Khan, Ahmed W. Mushtaha, Artificial intelligence in wastewater treatment: Research trends and future perspectives through bibliometric analysis, *Case Studies in Chemical and Environmental Engineering*, Volume 10, 2024, 100926, ISSN 2666-0164, <https://doi.org/10.1016/j.csee.2024.100926>.
2. Masud Parvez, Ahasanul Karim, Isa Ebtehaj, Hossein Bonakdari, Seddik Khalloufi, Artificial intelligence-assisted prediction of critical factors governing pesticide adsorption by agricultural wastes in wastewater treatment, *Journal of Environmental Chemical Engineering*, Volume 13, Issue 5, 2025, 118849, ISSN 2213-3437, <https://doi.org/10.1016/j.jece.2025.118849>.
3. Alali, Y., Harrou, F., & Sun, Y. (2023). Unlocking the Potential of Wastewater Treatment: Machine Learning Based Energy Consumption Prediction. *Water*, 15(13), 2349. <https://doi.org/10.3390/w15132349>
4. Molinos-Senante, M.; Maziotis, A. Evaluation of energy efficiency of wastewater treatment plants: The influence of the technology and aging factors. *Appl. Energy* 2022, 310, 118535.
5. El-Rawy, M.; Abd-Ellah, M.K.; Fathi, H.; Ahmed, A.K.A. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *J. Water Process Eng.* 2021, 44, 102380.
6. El Alaoui El Fels, A., Mandi, L., Kammoun, A., Ouazzani, N., Monga, O., & Hbid, M. L. (2023). Artificial Intelligence and Wastewater Treatment: A Global Scientific Perspective through Text Mining. *Water*, 15(19), 3487. <https://doi.org/10.3390/w15193487>
7. Andreides, M.; Dolejš, P.; Bartáček, J. The prediction of WWTP influent characteristics: Good practices and challenges. *J. Water Process Eng.* 2022, 49, 103009.
8. Yu, X., Chen, S., Zhang, X., Wu, H., Guo, Y., & Guan, J. (2023). Research progress of the artificial intelligence application in wastewater treatment during 2012-2022: a bibliometric analysis. *Water science and technology: a journal of the International Association on Water Pollution Research*, 88(7), 1750–1766. <https://doi.org/10.2166/wst.2023.296>
9. Zhang, S.; Wang, H.; Keller, A.A. Novel Machine Learning-Based Energy Consumption Model of Wastewater Treatment Plants. *ACS ES&T Water* 2021, 1, 2531–2540.
10. Sundui, B., Ramírez Calderón, O.A., Abdeldayem, O.M., Lázaro-Gil, J., Rene, E.R., & Sambuu, U. (2021). Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives. *Clean Technologies and Environmental Policy*, 23, 127-143.
11. Stefano Cairone, Shadi W. Hasan, Kwang-Ho Choo, Chi-Wang Li, Tiziano Zarra, Vincenzo Belgiorno, Vincenzo Naddeo, Integrating artificial intelligence modeling and membrane technologies for advanced wastewater treatment: Research progress and future perspectives, *Science of The Total Environment*, Volume 944, 2024, 173999, <https://doi.org/10.1016/j.scitotenv.2024.173999>.