

# Prediction of Dye Removal Using Machine Learning Techniques

Dilay Bozdağ Ak<sup>1</sup> , İhsan Hakan Selvi<sup>2</sup> 

<sup>1</sup> Sakarya University, Sakarya, Türkiye, [ror.org/04ttnw109](http://ror.org/04ttnw109)

<sup>2</sup> Sakarya University, Department of Information Systems Engineering, Sakarya, Türkiye, [ror.org/04ttnw109](http://ror.org/04ttnw109)

Corresponding author:

Dilay Bozdağ Ak,  
Department of Information Systems  
Engineering, Sakarya University  
[dilaybozdogak@sakarya.edu.tr](mailto:dilaybozdogak@sakarya.edu.tr)

Article History:

Received: 12.05.2025

Revised: 08.08.2025

Accepted: 11.08.2025

Published Online: 26.09.2025

## ABSTRACT

This study aims to predict the removal efficiency of methylene blue dye using experimental data collected from adsorption processes involving acorn-based biosorbents. A comparative evaluation of four machine learning algorithms (Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), Random Forest, and XGBoost) was conducted to determine the most suitable modeling approach. Two ANN architectures, with single and dual hidden layers respectively, achieved the highest predictive accuracy, with  $R^2$  values of 0.93 and 0.87. While XGBoost demonstrated better performance ( $R^2 = 0.64$ ) than Random Forest ( $R^2 = 0.61$ ), both ensemble models provided moderately accurate predictions. In contrast, the LSTM model performed poorly ( $R^2 = 0.44$ ), likely due to the non-sequential structure of the dataset. These findings underscore the potential of ANN-based models for accurately capturing nonlinear relationships in adsorption systems and also demonstrate the viability of alternative ensemble learning methods for predictive environmental modeling.

**Keywords:** Adsorption modeling, Artificial Neural Networks, XGBoost, LSTM, Dye removal, Predictive modeling

## 1. Introduction

Environmental pollution has become one of the most critical yet complex problems in recent years [1]. The rapid global population increase, accompanied by technological advancement and rapid industrialization, has significantly worsened the situation [2]. While industries contribute to human life by producing numerous products that meet basic needs, some of the waste generated during production disrupts ecosystems. In particular, the chemical waste released by manufacturers into the environment, whether directly or indirectly, has a detrimental impact on aquatic systems [3]. Unwanted dye waste and chemicals discharged during textile industry operations are a major contributor to environmental pollution. Water contamination is one of the outcomes of this pollution [4]. The large amounts of colored dyes and pigments discharged from industrial facilities pose a significant environmental threat. Dye waste, mostly from industrial and textile operations, has lethal and carcinogenic effects on aquatic life [5].

In recent years, artificial intelligence (AI)-based approaches have gained considerable attention for modeling complex environmental processes. Artificial Neural Networks (ANNs), inspired by the structure of the human brain, are one of the most widely used AI models. Composed of input, hidden, and output layers, ANNs are capable of learning nonlinear relationships in systems where conventional mathematical modeling falls short [6-7]. Their flexibility, generalization ability, and strong performance with limited data have made them especially popular in chemical and environmental engineering.

ANNs are computational models structured with an input layer, one or more hidden layers, and an output layer. These models are widely used to represent advanced systems and analyze the associations between input and output variables. Among the various types of ANNs available in literature, multi-layered architectures are particularly prevalent in engineering applications [9]. ANNs are especially advantageous when defining an explicit mathematical relationship for a given phenomenon is difficult or impossible. Thanks to their capability to capture nonlinear interactions among variables, ANNs are considered dependable, flexible, and practical tools across various disciplines [10]. As a result, the use of ANNs in modeling chemical and biochemical processes with intricate input-output dependencies has gained significant attention. In recent years, ANN-based approaches have been successfully applied in various processes, including dye removal, adsorption, fermentation, filtration, and drying.[11]. ANN provides valuable insights in dye removal studies using nonlinear regression data from experimental systems. In research involving ANN applied to adsorption systems, the percentage removal value is often selected as the prediction parameter to optimize the adsorption process [6].

A review of ANN applications in adsorption systems reveals several notable studies. Amouei et al. utilized sunflower seed powder as an adsorbent for cadmium (Cd) removal, examining the influence of various parameters, including pH, initial

concentration, and contact time. Their findings indicated that the ANN model closely matched the experimental results [12]. Similarly, S. M. El-Said et al. explored the removal of arsenite (As III) and arsenate (As V) from aqueous solutions using *Nigella sativa* L. (black cumin), applying ANN modeling to offer a new perspective on the data interpretation [13]. In another study, Garza-González et al. developed a genetic algorithm-optimized ANN model to remove methylene blue, incorporating three input variables, two hidden layers, and 20 neurons [14]. Çoruh et al. focused on removing malachite green and acid blue 161 dyes, analyzing the effects of dye concentration, temperature, and contact time. Their ANN model, which consisted of four input neurons, a hidden layer of 12 neurons, and an output neuron, showed a strong correlation with experimental observations [15]. Additionally, Öztürk et al. implemented an ANN model to predict methylene blue adsorption using waste sludge. The model, containing 12 neurons, effectively simulated the adsorption process [16].

Recent literature also highlights integration with novel adsorbents and machine learning models. Dey et al. (2024) employed artificial intelligence models, such as ANN and LSTM, to predict methyl blue dye removal using banana leaf-based adsorbents, reporting high accuracy and confirming the effectiveness of ANN-based approaches in adsorption modeling [17]. Kalsoom et al. (2023) demonstrated that the ultrasonic-assisted adsorption of pesticides onto activated carbon and biochar followed the Langmuir isotherm and pseudo-second-order kinetics, highlighting the efficiency of carbon-based materials for pollutant removal under optimized sonication conditions [18]. Ganthavee et al. (2024) optimized a three-dimensional electrochemical treatment process for dye removal using artificial neural network (ANN), support vector machine (SVM), and random forest models, reporting that ANN provided the highest predictive accuracy, with  $R^2$  values exceeding 0.90 for multiple performance parameters [19]. Bahrami et al. (2024) successfully employed random forest regression to model methylene blue adsorption onto polyethylene microplastics, achieving a high predictive accuracy ( $R^2 = 97.55\%$ ) and highlighting the significance of initial dye concentration and adsorbent dose in determining adsorption performance [20]. Yaseen and Alhalimi (2025) applied various ensemble machine learning models, including XGBoost and Random Forest, to predict heavy metal adsorption onto biochar, reporting the highest predictive accuracy ( $R^2 = 0.921$ ) with XGBoost and emphasizing pH, pyrolysis temperature, and initial concentration ratio as key influencing factors [21].

While these studies collectively illustrate the versatility of AI-based models in adsorption processes, the present study aims to evaluate the performance of four different machine learning algorithms comparatively—Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), Random Forest (RF), and XGBoost—in predicting the removal efficiency of methylene blue using a sustainable and low-cost adsorbent: acorn (oak nut) powder. Experimental data from batch adsorption tests were used to train and validate the models. The results revealed that ANN models exhibited superior predictive performance, though ensemble methods like XGBoost also provided competitive accuracy, highlighting the potential of hybrid AI approaches in environmental modeling.

## 2. Method

### 2.1. Dataset Description

The dataset comprises 76 experimental observations collected under varying conditions of pH, temperature, contact time, adsorbent dosage, and initial dye concentration. This study utilized 76 data points, with 60 assigned to training, 8 to validation, and 8 to testing. In regression analysis, the goal is to predict the value of the output variable based on the values of the input variables.

### 2.2. Data Preprocessing and Performance Evaluation

Before model training, all input features were normalized using the min-max normalization technique to scale values between 0 and 1. This ensured that each feature contributed equally during the learning process and prevented dominance by attributes with larger numerical ranges.

After model training, the performance of the predictive models was evaluated by comparing the predicted values to the actual experimental results. As the expected values do not exactly match the ones observed, statistical error metrics were applied to quantify the deviation. These metrics include Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination ( $R^2$ ).

In regression problems, the ideal regression curve is the one that best fits the actual data points. The deviation of each data point from this curve is minimized using the least squares method, which aims to reduce the total squared error.

The MSE is calculated using Equation (1):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y - y')^2 \quad (1)$$

Where:

$n$  = number of data points

$y$  = observed values

$y'$  = predicted values

The MAPE equation is shown in Equation (2):

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \tag{2}$$

Where:

n = number of summation iterations

A<sub>t</sub> = actual value

F<sub>t</sub> = predicted value

The MAE equation is presented in Equation (3):

$$MAE = \frac{\sum_{t=1}^n |y_t - x_t|}{n} \tag{3}$$

n = total number of data points

y<sub>t</sub> = actual value

x<sub>t</sub> = predicted value

### 2.3. Model Architectures

The overall modeling workflow followed in this study is illustrated in Figure 1. This includes data preprocessing, model selection, training, and performance evaluation steps applied across four different machine learning algorithms: ANN, LSTM, Random Forest, and XGBoost.

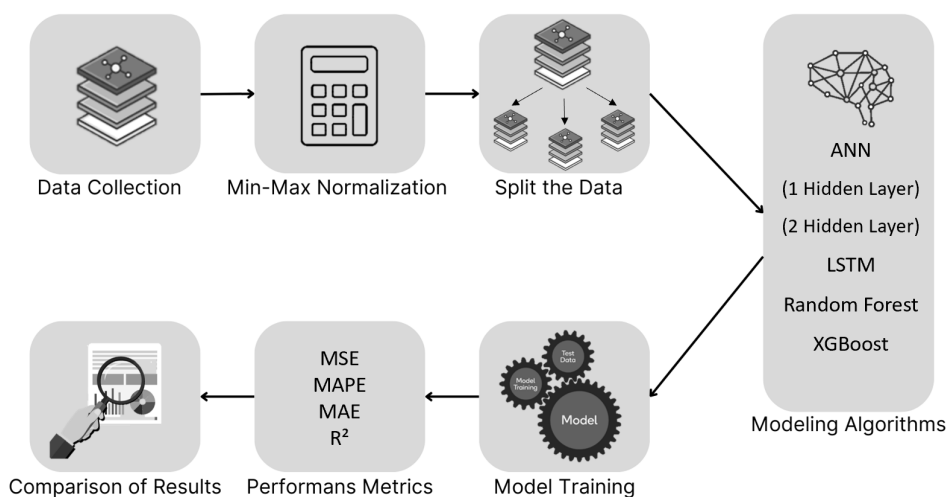


Figure 1. Workflow of the modeling process used in the study.

#### 2.3.1. Modeling Adsorption Using ANNs

In this study, the removal of methylene blue dye using an acorn-based adsorbent was modeled using Artificial Neural Networks (ANNs). The input parameters included solution pH, temperature, adsorbent dosage, contact time, and initial dye concentration, while the output parameter was the percentage removal. The ANN architecture consisted of three main layers: an input layer, one or more hidden layers, and an output layer. Each neuron in a layer was fully connected to neurons in the subsequent layer, with no intra-layer connections.

##### 2.3.1.1. ANN Model with a Single Hidden Layer

In the first ANN architecture, a single hidden layer was used. The Levenberg-Marquardt algorithm (LMA) was employed for training this model. A trial-and-error approach was followed to determine the optimal number of neurons in the hidden layer. The best performance was obtained with 12 neurons, achieving high prediction accuracy with minimized MSE. This configuration successfully modeled the nonlinear relationship between the adsorption parameters and dye removal efficiency. The general structure of the model is illustrated in Figure 2.

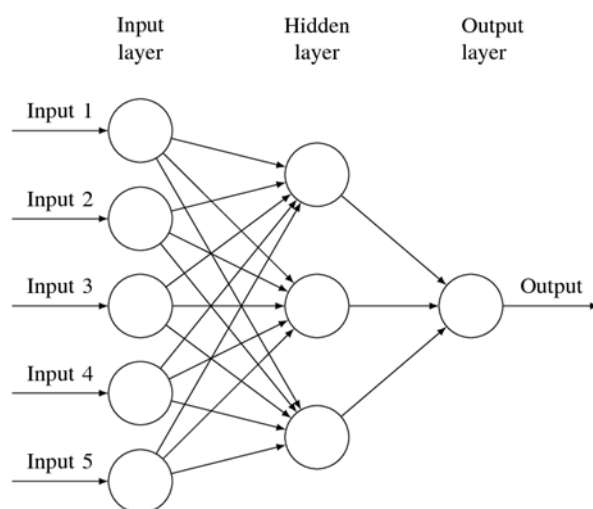


Figure 2. Structural diagram of the ANN model

### 2.3.1.2. ANN Model with Two Hidden Layers

In the second ANN configuration, a deeper architecture was implemented, incorporating two hidden layers with 24 and 12 neurons, respectively. The ReLU activation function was used in the hidden layers, and a linear activation function was applied in the output layer. The model was trained over 1000 epochs using the Adam optimizer, with early stopping applied to prevent overfitting. This deeper network demonstrated strong generalization capability and yielded a lower MAPE and MAE on the test dataset, confirming the feasibility of using deeper ANN structures in dye adsorption modeling.

Both ANN architectures exhibited high predictive power, with slight variations in performance depending on the depth and training dynamics. The inclusion of both shallow and deep models allowed a comparative evaluation of ANN performance in capturing the nonlinearities of the adsorption system.

### 2.3.2. Modeling with LSTM (Long Short-Term Memory)

To evaluate the capabilities of deep learning methods, an LSTM-based (Long Short-Term Memory) model was also developed. LSTM, a type of Recurrent Neural Network (RNN), is particularly suited for modeling temporal dependencies in sequential data. The model used a stacked LSTM architecture with dropout regularization to mitigate overfitting and was trained using an 80:20 train-test split. Despite optimizing hyperparameters and employing early stopping, the model exhibited limited predictive performance due to the non-sequential structure of the dataset.

### 2.3.3. Modeling with Random Forest Regressor

A Random Forest Regressor (RFR) was implemented as an ensemble learning approach based on decision trees. It constructs multiple trees on randomly sampled subsets of the dataset and averages the outputs for prediction. This technique reduces variance and improves generalization. The model was trained using bootstrap aggregation (bagging), and performance was evaluated on the test data using MSE, MAE, and  $R^2$  metrics.

### 2.3.4. Modeling with XGBoost Regressor

XGBoost (Extreme Gradient Boosting) was applied to enhance prediction accuracy further. XGBoost builds additive decision trees in a forward, stage-wise manner, utilizing gradient boosting optimization. This method is known for its speed and performance on structured data. The model's hyperparameters were tuned, and its predictions were compared with the actual dye removal percentages using standard evaluation metrics.

## 3. Results and Discussion

### 3.1. ANN Modeling Study with One Hidden Layer

In this part of the study, the performance of the ANN model with a single hidden layer was evaluated in predicting the percentage removal of methylene blue. The optimal network architecture was determined by experimenting with different numbers of neurons, where the best results were obtained using 12 hidden neurons, trained with the Levenberg–Marquardt algorithm (LMA).

The general structure of the network, including five input variables (solution pH, adsorbent dosage, contact time, temperature, and initial concentration), one hidden layer, and one output neuron, is presented in Figure 3.

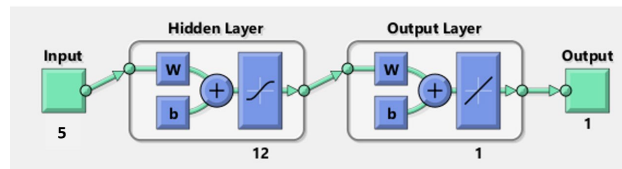


Figure 3. Number of inputs, hidden, and output layers

The model was trained using standard backpropagation, and the best performance was observed at the 30th epoch, as seen in the MSE convergence graph (Figure 4). This indicates a stable learning curve during training, validation, and testing phases.

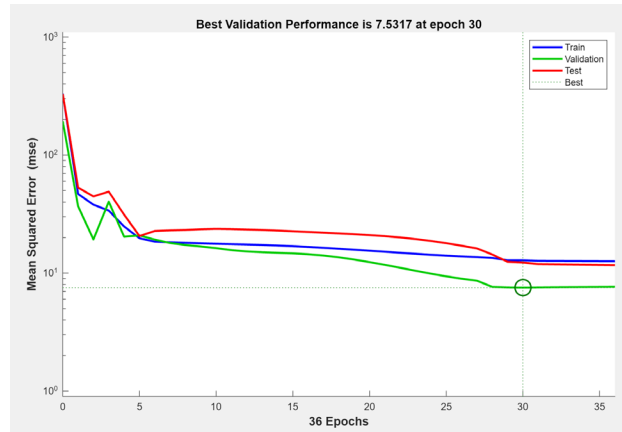


Figure 4. Error Convergence of Training, Validation, and Test Data with Best Performance Highlighted

The comparison of actual and predicted values is shown in Figure 5, where the expected values closely follow the experimental data, particularly in the test dataset. Figure 5 displays the predicted normalized removal values generated by the ANN model for the training and test datasets, plotted against the experimentally obtained normalized values. The figure reveals a strong alignment between the predicted and actual results. This indicates that the ANN model's predictions agree with the experimental data, demonstrating that the developed model effectively captures the behavior observed in the analyses.

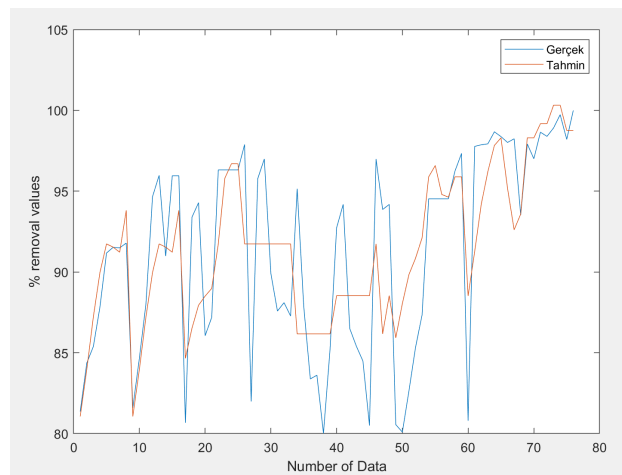


Figure 5. Comparison of predicted and actual percentage removal outputs

To further assess model quality, error distribution across all data subsets was analyzed. As illustrated in Figure 6, most prediction errors are clustered around zero, indicating low deviation and high consistency.

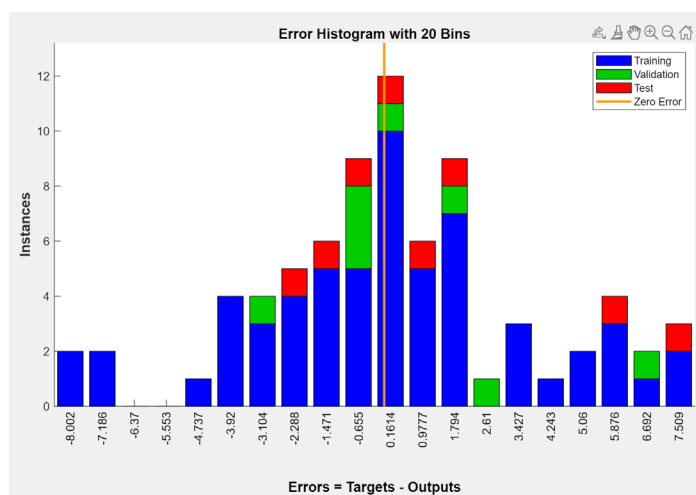


Figure 6. Histogram Graph

Training, validation, and test data are represented by blue, green, and red bars, respectively. The histogram highlights outlier data points associated with lower performance. The error histogram is presented in Figure 6, which illustrates the distribution of prediction errors (calculated as Errors = Targets - Outputs) across the training, validation, and test datasets, segmented into 20 bins. It can be seen that most errors are clustered around zero, indicating a strong agreement between the model's predictions and the actual values. This clustering near zero error demonstrates the model's high predictive accuracy and effective learning capability. In conclusion, the error histogram confirms that the developed model demonstrates reliable performance, with low prediction errors predominantly centered around zero, and maintains consistency across all data subsets. This supports the validity and robustness of the model for predictive tasks within the studied system.

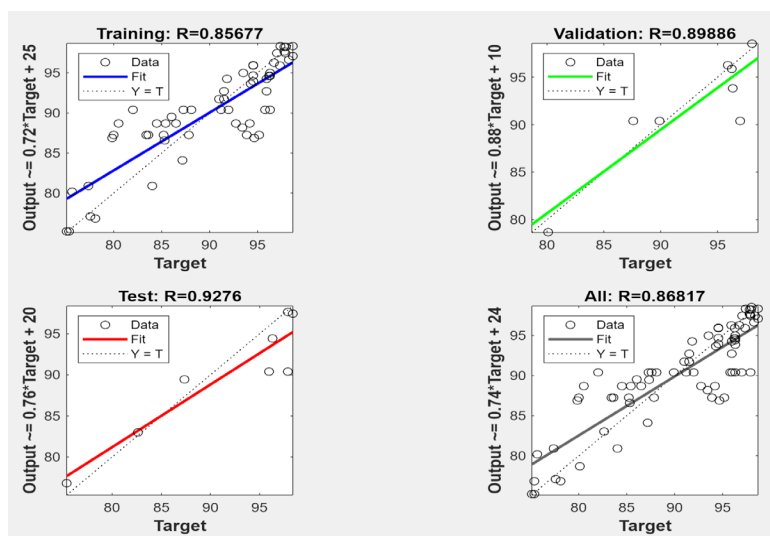


Figure 7. Regression analysis is performed on training, validation, test, and all data sets.

As shown in Figure 7(a), the ANN model correlates well with the training data for the optimized structure. Despite some scatter in the data, it also demonstrated relatively good performance in validation and testing, as illustrated in Figures 7(b) and 7(c). Overall, the ANN model exhibited satisfactory prediction performance for the batch adsorption experimental dataset, as shown in Figure 7(d).

Figure 6 presents the ANN model. The MSE values for the training, validation, and test sets were calculated as  $12.8271e-0$ ,  $7.5317e-0$ , and  $12.2880e-0$ , respectively. The associated  $R^2$  values were 0.8568, 0.8989, and 0.9276, demonstrating strong predictive accuracy, particularly in the validation and test datasets.

In this study, dye removal was evaluated in terms of percentage removal using both ANN and regression analysis. The experimental and predicted values were statistically assessed using the MAPE method.

Table 1 presents the data points allocated for training, validation, and testing, along with each dataset's MSE and  $R^2$  values.

Table 1. Performance evaluation metrics for the single-hidden-layer ANN model

	Observations	MSE	R <sup>2</sup>
Training	60	12.8271	0.8568
Validation	8	7.5317	0.8989
Test	8	12.2880	0.9276

Using Equation (2), the MAPE value was calculated as 0.0418. A MAPE value below 10% indicates that the model is highly successful. The R<sup>2</sup> value, commonly used in AI studies, is displayed in graphs that show experimental and predicted data. R<sup>2</sup> is a statistical method used to measure the degree of relationship between two or more variables or the linear relationship between two variables. The number of neurons in the hidden layer was determined through a trial-and-error method based on the obtained R<sup>2</sup> values. The overall R<sup>2</sup> value was found to be 87%. Due to their toxic effects, dyes pose a significant threat, especially in aquatic systems. Traditional experimental methods for dye removal often result in the release of additional chemicals into the environment. Reducing the number of experiments through AI studies is crucial in minimizing chemical pollution. AI and ANNs are increasingly widespread today, as they allow for the simulation of experimental outcomes in complex systems through virtual laboratories. The main objective here is to estimate approximate values for multiple parameters using AI and experimental results. Consequently, future studies can benefit from time savings, a reduction in the number of experiments, and decreased experimental costs. Conducting fewer experiments results in less chemical usage and a tangible reduction in environmental pollution. Additionally, since setting up laboratories and devices for such experiments involves high costs, this study also reduces those expenses. The close alignment of AI and ANNs with actual experimental outcomes is crucial for sustainability. Being environmentally friendly and cost-effective are key indicators of sustainability. Furthermore, the fact that the real experimental results closely match the predictions of AI and ANN models demonstrates that this study is feasible from an applicability perspective.

### 3.2. ANN Modeling Study with Two Hidden Layers

In addition to the single-hidden-layer ANN modeling results, an ANN prediction model was also developed using a two-hidden-layer model. The modeling process was carried out using Python's TensorFlow and scikit-learn libraries. For this purpose, the dataset of 76 samples was divided into three parts: 60 for training, 8 for validation, and 8 for testing. Figure 8 illustrates the model architecture, which consists of 5 input neurons, two hidden layers (with 12 and 24 neurons, respectively), and one output neuron.

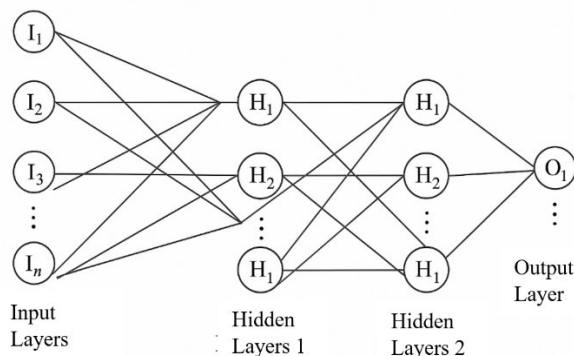


Figure 8. Model Architecture

The hidden layers used the ReLU activation function, while the output layer used a linear activation function. As shown in Figure 8, the training process was carried out over 1000 epochs, and early stopping was applied to prevent the model from overfitting.

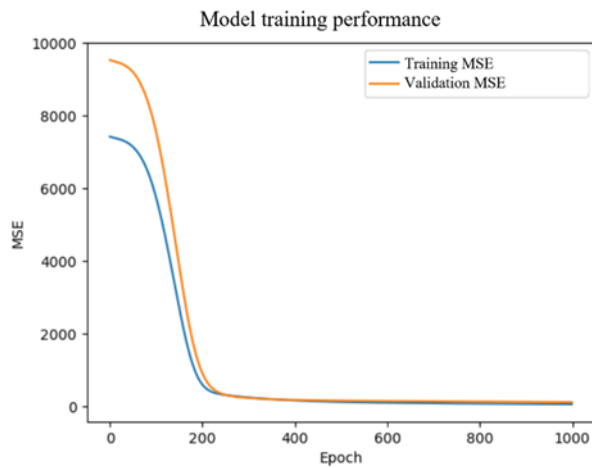


Figure 9. Model Training Performance

Figure 9 illustrates that the training and validation MSE values decreased steadily throughout the epoch, indicating a stable learning curve for the model.

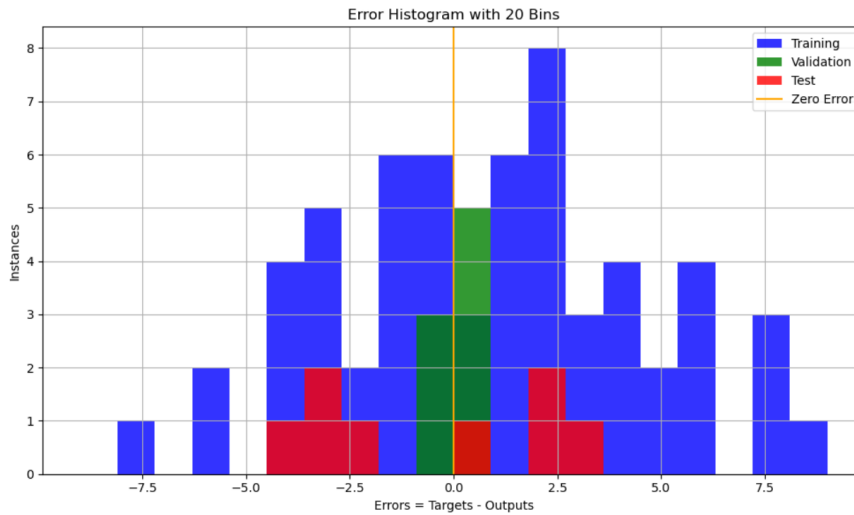


Figure 10. Histogram Graph

Training, validation, and test data are represented by blue, green, and red bars, respectively. The histogram presents the distribution of prediction errors (calculated as  $Errors = Targets - Outputs$ ) across all three datasets. As shown in Figure 10, the majority of errors are densely concentrated around zero, indicating that the predicted values closely match the actual values. This tight clustering around zero signifies the model’s strong learning capability and high prediction accuracy. Additionally, the limited number of data points with significant errors supports the consistency of the model across different subsets. The inclusion of the orange vertical line at zero error visually emphasizes the symmetry and balance of the error distribution. These results validate the reliability and robustness of the developed model in providing accurate predictions across training, validation, and test datasets.

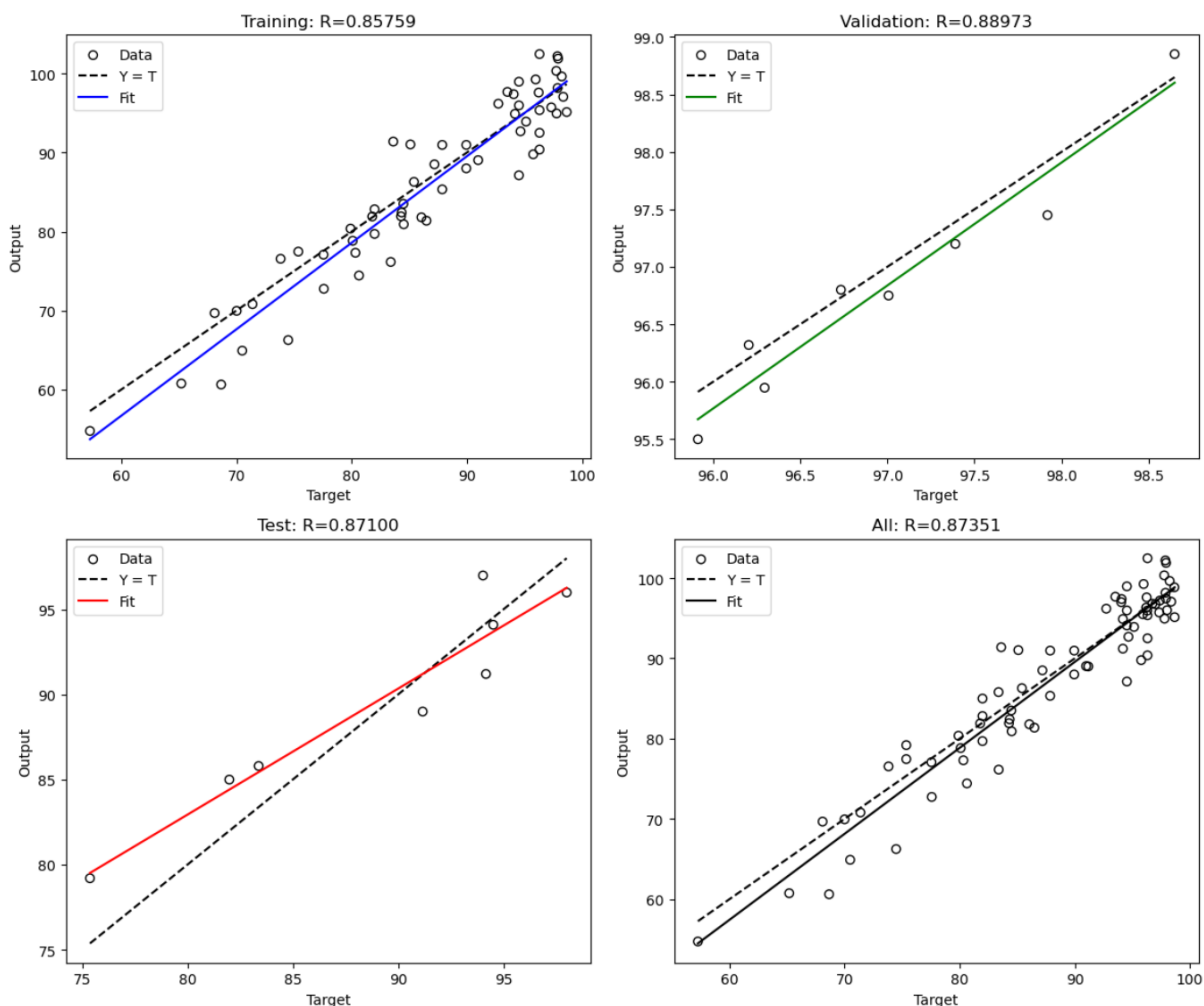


Figure 11. Regression analysis is performed on training, validation, test, and all data sets.

As illustrated in Figure 11(a), the regression analysis of the training data reveals a strong correlation between the predicted and actual values, with a coefficient of determination ( $R^2$ ) of 0.85759, confirming that the ANN model effectively captures the underlying patterns in the training set. Figure 11(b) further demonstrates good agreement with the validation data, with an  $R^2$  value of 0.88973, indicating the model’s ability to generalize well to unseen data. Similarly, the test data in Figure 11(c) yields an  $R^2$  value of 0.87100, reinforcing the robustness and predictive accuracy of the model. The combined regression plot in Figure 11(d) summarizes the performance across all datasets, yielding an overall  $R^2$  of 0.87351. This comprehensive performance evaluation confirms the ANN model’s reliability and suitability for the predictive modeling task within the experimental framework.

The model performance was evaluated using the following statistical metrics:

MSE: 9.05

MAE: 2.86

MAPE: 2.96%

The model’s absolute and relative error values are pretty low. A MAPE value below 3% indicates that the developed ANN model provides highly accurate predictions for dye removal. These metrics demonstrate that the model performs effectively on the training and previously unseen test data.

Table 2. Performance evaluation metrics for the two-hidden-layer ANN model

	Observations	MSE	$R^2$
<b>Training</b>	60	11.1254	0.8576
<b>Validation</b>	8	8.7856	0.8897
<b>Test</b>	8	9.0500	0.8710

Examining the training and validation error graphs reveals that the model follows a stable learning curve, with the validation error closely tracking the training error. The scatter plot comparing the actual and predicted percentage removal values shows that the predictions are closely aligned with the reference line, indicating high accuracy.

### 3.3. Modeling with LSTM (Long Short-Term Memory)

An LSTM (Long Short-Term Memory)-based model was implemented using the same dataset to evaluate the predictive capacity of deep learning architecture. LSTM is a special Recurrent Neural Network (RNN) capable of learning long-term dependencies and is frequently used in time-series prediction tasks.

The model consisted of a stacked LSTM architecture with multiple layers, incorporating dropout regularization to mitigate overfitting. It was trained using a dataset split into 80% training and 20% testing. Despite applying early stopping and tuning parameters such as the number of epochs, batch size, and neuron count, the LSTM model failed to deliver satisfactory results.

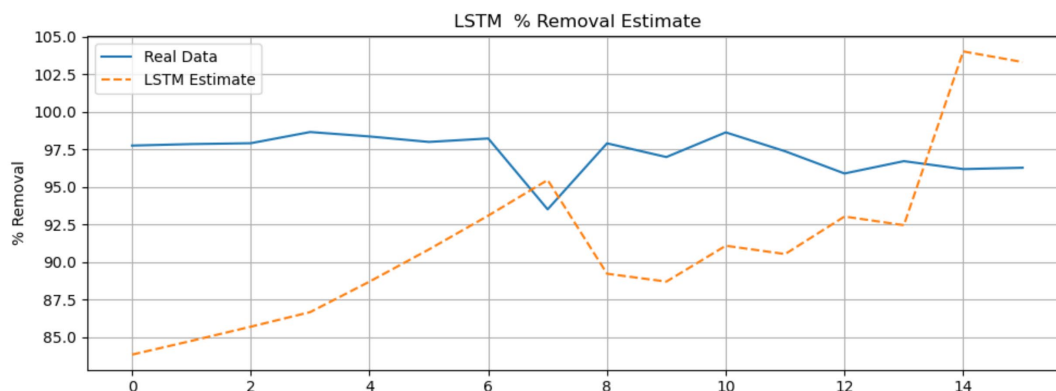


Figure 12. LSTM Real Estimate Performance

Figure 12 compares actual and predicted values. Although some values were predicted reasonably well, the general performance remained suboptimal. The predicted values exhibited a high deviation from the exact values, especially in the mid-range data.

The model performance metrics were as follows:

MSE: 27.93

MAE: 4.51

R<sup>2</sup>: 0.44

The dataset's absence of a sequential or time-series structure can explain the LSTM model's limited performance. Since LSTM networks are specifically designed to capture patterns across ordered temporal data, their applicability becomes constrained when such temporal dependencies are absent.

Although LSTM is a powerful deep learning model in domains with temporal patterns, such as speech recognition, weather forecasting, or stock prediction, it is not suitable for this dataset. The lack of inherent time dependencies limited LSTM's ability to capture meaningful patterns. ANN models, particularly those using the Levenberg-Marquardt algorithm and TensorFlow-based architectures, were better suited to the dataset and consistently outperformed LSTM in all evaluated metrics.

### 3.4. Modeling with Random Forest Regressor

The Random Forest Regressor (RFR) was implemented to explore the applicability of ensemble learning methods in modeling nonlinear adsorption systems. Random Forest is a decision tree-based ensemble algorithm that reduces variance and improves generalization through bootstrap aggregation (bagging). Its robustness and interpretability make it an attractive candidate for engineering problems with complex input-output relationships.

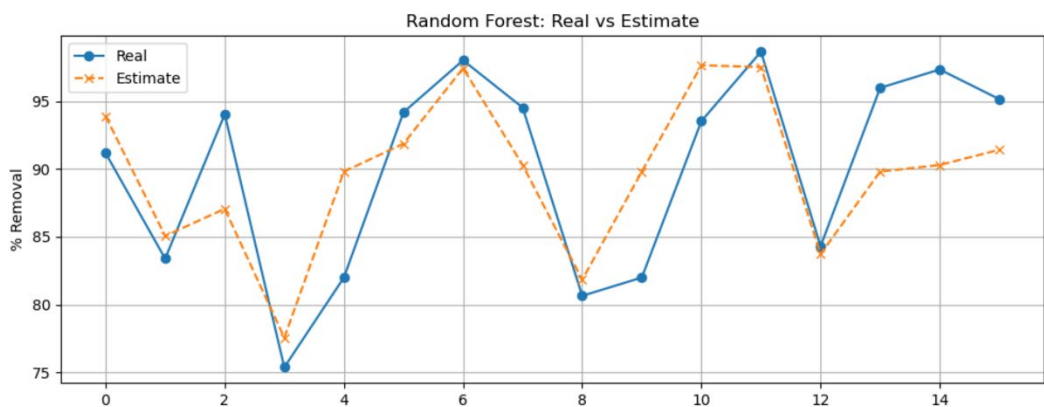


Figure 13. Random Forest Real Estimate Performance

Figure 13 summarizes the model’s prediction performance on the test set. It shows the actual versus predicted dye removal percentages, with the prediction trend closely following the experimental values.

Table 3. Feature importance scores of input variables derived from the Random Forest model.

Input Variable	Importance Score
Initial Dye Concentration	0.51
Adsorbent Dosage	0.21
Contact Time	0.16
pH	0.11
Tempature	0.03

Table 3 presents the feature importance scores derived from the Random Forest model. The results indicate that the initial dye concentration had the most significant influence on the model output, followed by the adsorbent dose and contact time. pH and temperature contributed to a lesser extent. These findings align with the known behavior of dye adsorption systems, where concentration gradients play a dominant role in removal efficiency.

The statistical evaluation metrics for the Random Forest model were:

MSE: 20.67

MAE: 3.77

R<sup>2</sup>: 0.61

Although it did not outperform the ANN-based models, the RFR model demonstrated a reasonable generalization capability. Its slightly lower accuracy is expected due to its non-parametric, tree-based nature and the lack of internal optimization mechanisms, such as backpropagation in neural networks.

### 3.5. Modeling with XGBoost Regressor

The XGBoost Regressor (XGBRegressor) was implemented to enhance prediction accuracy further and evaluate gradient-boosted ensemble methods. XGBoost (Extreme Gradient Boosting) is a robust machine learning algorithm based on gradient boosting of decision trees. Its advantages include high efficiency, scalability, and superior performance on structured/tabular data, which makes it a strong candidate for modeling nonlinear adsorption systems.

Figure 14 shows the comparison between predicted and actual values. The model demonstrates a more substantial alignment with experimental data compared to previous tree-based methods, such as Random Forest.

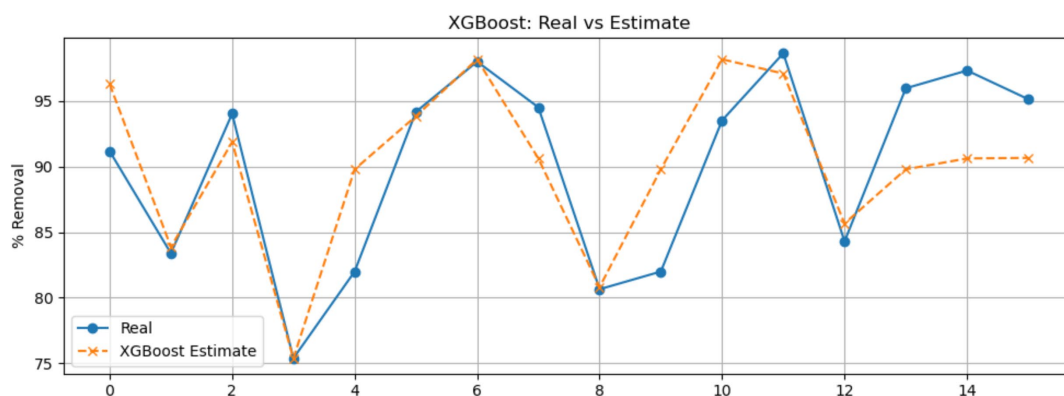


Figure 14. XGBoost Real Estimate Performance

The model performance was evaluated using standard statistical indicators:

MSE: 18.69

MAE: 3.28

R<sup>2</sup>: 0.64

These values indicate a moderate improvement over the Random Forest Regressor, suggesting that XGBoost has captured a larger portion of the dataset's nonlinear variance.

#### 4. Conclusion

In this study, the percentage removal efficiency of methylene blue dye using an acorn-based adsorbent was modeled and evaluated using various machine learning and statistical approaches, including ANNs, LSTM, Random Forest Regressor, and XGBoost Regressor. Among all models tested, the ANN architecture provided the most accurate and reliable predictions. The ANN model with a single hidden layer trained using the Levenberg-Marquardt algorithm achieved the highest performance, with an R<sup>2</sup> of 0.93. The deeper ANN model, composed of two hidden layers trained with early stopping and ReLU activation functions, also yielded strong results with an R<sup>2</sup> of 0.87.

In contrast, the LSTM model exhibited the weakest performance due to the non-sequential nature of the dataset, which limited its ability to capture meaningful temporal patterns. While ensemble-based methods such as Random Forest and XGBoost demonstrated moderate predictive capability, they did not match the accuracy achieved by the ANN models. XGBoost, benefiting from gradient boosting, slightly outperformed Random Forest.

Table 3. Comparison with ANN and Other Models

Model	MSE	MAE	R <sup>2</sup>
Two Hidden Layer ANN	9.05	2.86	0.87
One Hidden Layer ANN	12.29	2.86	0.93
XGBoost	18.69	3.28	0.64
Random Forest	20.67	3.77	0.61
LSTM	27.93	4.51	0.44

It demonstrates the superiority of ANN models in capturing the complex, nonlinear behavior of the adsorption system, as shown in Table 3. ANNs provide highly accurate predictions with low error criteria, confirming their suitability for modeling environmental processes.

These results demonstrate the superiority of ANN models in capturing the nonlinear and complex behavior of adsorption systems. In particular, ANNs achieved highly accurate predictions with low error metrics, confirming their suitability for modeling environmental processes. Moreover, the successful implementation of a two-hidden-layer ANN using open-source tools, such as Python, underscores the practicality and accessibility of such platforms for scientific modeling. The application of ANN in this study not only reduced the need for extensive experimentation but also provided benefits in terms of cost reduction and environmental sustainability.

Overall, this study contributes to existing literature by providing a comparative analysis of multiple machine learning methods for modeling dye adsorption. It reinforces the effectiveness of ANN-based approaches in non-temporal datasets and underlines their potential to minimize experimental workload, costs, and environmental impact in future adsorption research.

## References

- [1] A. Asfaram *et al.*, "Preparation and characterization of MnO. 4ZnO. 6Fe<sub>2</sub>O<sub>4</sub> nanoparticles supported on dead cells of *Yarrowia lipolytica* as a novel and efficient adsorbent/biosorbent composite for the removal of azo food dyes: central composite design optimization study," *ACS Sustainable Chem. Eng.*, vol. 6, no. 4, pp. 4549-4563, 2018.
- [2] M.R. Gaddekar and M.M. Ahammed, "Modelling dye removal by adsorption onto water treatment residuals using combined response surface methodology-artificial neural network approach", *J. Environ. Manag.*, vol. 231, pp. 241–248, 2019..
- [3] A.S. Al-Wasidi, F.A. Saad, S. AlReshaidan and A.M. Naglah, "Facile Synthesis of ZSM-5/TiO<sub>2</sub>/Ni Novel Nanocomposite for the Efficient Photocatalytic Degradation of Methylene Blue Dye", *J. Inorg. Organomet. Polym. Mat.*, vol. 32, pp. 3040–3052, 2022.
- [4] S. Vajnhandl and J.V. Valh, "The status of water reuse in European textile sector", *J. Environ. Manage.*, vol. 141, pp. 29-35, 2014.
- [5] M. F. Pinheiro, G.S. Rodrigues, J.A. Junior, R. de Sousa, and de A.R. da Costa, "Analysis of the adsorptive capacity of arabic coffee straw using blue methylene dye", *Braz. J. Dev.*, vol. 6, no. 1, pp. 2861-2868, 2020.
- [6] A. Arı, ve M.E. Berberler, "Yapay Sinir Ağları ile Tahmin ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı", *Acta Infologica*, vol.1, no.2, pp. 55-75, 2017.
- [7] H. Esen, "Düşey borulu toprak kaynaklı ısı pompasının konutlardaki iklimlendirme sistemlerinde mevsimsel davranışın araştırılması", Doktora Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ 2007.
- [8] T. Partal, "Türkiye yağış miktarının yapay sinir ağları ve dalgacık dönüşümü yöntemleri ile tahmini", Doktora Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2007.
- [9] N. G. Polsona, and V.O. Sokolov, "Deep learning for short-term traffic flow prediction, Transportation Research Part C:" *Emerging Technologies*, vol.79, pp.1-17, 2017
- [10] A. Kardam, K.R. Raj, J. K. Arora and S. Srivastava, "Simulation and Optimization of Artificial Neural Network Modeling for Prediction of Sorption Efficiency of Nanocellulose Fibers for Removal of Cd (II) Ions from Aqueous System", *Eng. Phys. Sci.*, vol. 11, no. 6, pp. 497-508, 2014.
- [11] N. Donut, and L. Cavas, "Artificial Neural Network Modeling of Tetracycline Biosorption by Pre-treated *Posidonia oceanica*", *Turkish J. Fish. Aquat. Sc.*, vol.17, pp. 1317-1333, 2017.
- [12] A.A. Amouei, F. Amooey and F. Asgharzadeh, "A study of cadmium removal from aqueous solutions by sunflower powders and its modeling using artificial neural network", *Iran. J. Health Sci.*, vol. 1, no.3, pp. 28-34, 2013.
- [13] S. M. El-Said *et al.*, "Adsorptive removal of Arsenite as (III) and Arsenate as (V) heavy metals from wastewater using *Nigella sativa* L.", *J. Asian Sci. Res.*, vol. 2, pp. 96-104, 2009.
- [14] M. Garza-González *et al.*, "Artificial neural network for predicting biosorption of methylene blue by *Spirulina* sp.", *Water Sci. Technol.*, vol. 75, no.5, pp. 977-983, 2011.
- [15] S. E. Çoruh, E. Kılıç and F. Geyikci, "Prediction of adsorption efficiency for the removal of malachite green and acid blue 161 dyes by waste marble dust using ANN", *Global Nest J.*, vol. 16, no.4, pp. 676-689, 2014.
- [16] N. Öztürk, H.B. Şentürk, A. Gündoğdu and C. Duran, "İçme suyu arıtma tesisi atık çamuru üzerine metilen mavisi adsorpsiyonu ve yapay sinir ağları ile modellenmesi", *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, vol. 25, no. 2, pp. 1083-1103, 2020.
- [17] Barna, S. D., Jahan, M. N., Sium, S. R., Nag, A., Ali, M. H., & Dutta, S. K. (2024). Sustainable cost-effective chemically modified banana leaf powder for methyl blue dye removal: kinetics, isotherm, thermodynamics and artificial intelligence-based analysis. *Discover Chemistry*, 1(1), 38.
- [18] Kalsoom, Ali, A., Khan, S., Ali, N., & Khan, M. A. (2024). Enhanced ultrasonic adsorption of pesticides onto the optimized surface area of activated carbon and biochar: adsorption isotherm, kinetics, and thermodynamics. *Biomass Conversion and Biorefinery*, vol. 14, no. 14, pp. 15519-15534.
- [19] Ganthavee, V., Fernando, M. M., & Trzcinski, A. P. (2024). Monte carlo simulation, artificial intelligence and machine learning-based modelling and optimization of three-dimensional electrochemical treatment of Xenobiotic Dye wastewater. *Environmental Processes*, vol. 11, no. 3, p. 41.
- [20] Bahrami, M., Amiri, M. J., Rajabi, S., & Mahmoudi, M. (2024). The removal of methylene blue from aqueous solutions by polyethylene microplastics: Modeling batch adsorption using random forest regression. *Alexandria Engineering Journal*, 95, pp. 101-113.

- [21] Yaseen, Z. M., & Alhalimi, F. L. (2025). Heavy metal adsorption efficiency prediction using biochar properties: a comparative analysis for ensemble machine learning models. *Scientific Reports*, vol. 15, no. 1, p. 13434.

### **Article Information Form**

#### **Authors Contributions**

The artificial intelligence studies, code development, literature review, and the writing of the manuscript were carried out by Dilay BOZDAĞ AK. Hakan İhsan SELVİ reviewed the manuscript.

#### **Acknowledgments**

We thank Prof. Dr. Esra ALTINTIĞ for providing the study data.

#### **Conflict of Interest Notice**

The authors declare that no conflicts of interest are associated with this article's publication.

#### **Ethical Approval**

It is affirmed that the preparation of this study was conducted under scientific and ethical standards, and all referenced sources have been duly cited in the bibliography.

#### **Availability of data and material**

Not applicable / or link

#### **Artificial Intelligence Statement**

The authors confirm that no artificial intelligence tools were employed in the preparation or authorship of this article.

#### **Plagiarism Statement**

This article has been scanned by iThenticate™.