Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7, 536-549
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i7.8654
© 2025 by the authors; licensee Learning Gate

Efficient AI-based water quality prediction and classification for sustainable urban environments in Texas city

Shubham Kuppili[1], Victor Sheng[2], Kishore Kunal[3], R. Murgadoss[4], Vairavel Madeshwaren[5*]
[1,2]Department of Computer Science, Texas Tech University, Lubbock, Texas, USA.
[3]Loyola Institute of Business Administration, Chennai, Tamil Nadu, India.
[4]Department of Artificial Intelligence and Data science, VSB College of Engineering & Technical Campus, Coimbatore, Tamil Nadu, India.
[5]Department of Agriculture engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamil Nadu, India; phdannauniv2020@gmail.com (V.M.).

Abstract: This study utilizes advanced technologies to address water contamination by analyzing data collected from the Red River, Rio Grande, and Trinity River in Texas City, USA, between March 2023 and March 2024. The dataset comprises seven critical water quality parameters—conductivity, pH, turbidity, dissolved oxygen (DO), total and fecal coliform, chemical oxygen demand (COD), and nitrate—ensuring compliance with U.S. government standards for safe and clean drinking water. The proposed Efficient AI-based Water Quality Prediction and Classification (EAI-WQP) model aims to accurately predict water pollution parameters, particularly those influenced by industrial activities. Leveraging Apache Spark, a powerful big data processing framework, the model enables real-time data handling and analysis for effective pollution management. To enhance prediction accuracy, the model's parameter tuning is optimized using the Firefly Algorithm (FA). Furthermore, an Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier is integrated into the model, combining fuzzy logic with neural networks to classify water quality into pollutant and non-pollutant categories. Comparative evaluations against established machine learning techniques such as GRU-ARIMA, SVM, and Random Forest demonstrate the superior performance of the EAI-WQP model in terms of accuracy, precision, F1 score, and recall. The study aims to analyze water pollution in Texas City using advanced AI methodologies, develop the EAI-WQP model for accurate forecasting of water quality parameters, implement real-time big data processing with Apache Spark, optimize model performance using the Firefly Algorithm (FA), classify water quality using ANFIS, and demonstrate the model's superiority over traditional machine learning methods.

Keywords: *Environmental management, GRU-ARIMA, machine learning, Industrial pollution, Water quality prediction.*

1. Introduction

The main aim of the study was to develop and evaluate the Efficient AI-based Water Quality Prediction and Classification (EAI-WQP) model to accurately forecast water pollution parameters. By leveraging technological advancements such as Apache Spark for real-time data processing and the Firefly Algorithm for parameter tuning, the study sought to enhance prediction accuracy. The model, which incorporates the Adaptive Neuro-Fuzzy Inference System (ANFIS), was designed to improve water quality classification. Additionally, the study aimed to compare the performance of the EAI-WQP model with other machine learning techniques, including GRU-ARIMA, SVM, and Random Forest, to identify the most effective approach for predicting water pollution levels.

The environment, human health, and water resources all greatly depend on water quality (WQ). To address this worldwide challenge, practitioners and researchers have gotten heavily involved in water quality monitoring and modeling. This is because billions of people on Earth depend on clean, healthy,

History: Received: 3 April 2025; Revised: 5 June 2025; Accepted: 10 June 2025; Published: 7 July 2025
* Correspondence: phdannauniv2020@gmail.com

and sufficient freshwater. Water quality (WQ) can be estimated and the degree of contamination can be determined by utilizing a combination of the water's physical, chemical, and biological aspects. Because there have been so many instances of water contamination in recent years, environmental management organizations around the world have become increasingly interested in the assessment and estimation of WQ. The environmental infrastructure depends on specific case study surface water quality studies. The most important markers of WQ are dissolved oxygen (DO) and biochemical oxygen demand (BOD) because they affect a variety of biological, chemical, and physical characteristics of water. Controlling stream pollution, managing river water quality, and supporting natural processes all depend on the accurate evaluation of these two variables. These quality variables are still determined using traditional techniques (volumetric titration), which are more subjective than instrumental techniques. Much money, time, and effort could be saved if these WQ elements could be predicted with a reasonable degree of precision. Because of this, researchers have developed reliable models to forecast BOD and DO from other easily accessible inputs related to water quality.

However, the prediction and quantitative modeling of surface water quality variables have proven to be a challenging task in recent decades. DO and BOD are influenced by a wide range of biotic and abiotic factors, as well as the complex relationships between them. Most of these interactions are still vague and undefinable at this point, and it is difficult to get the data needed for process modeling. As a result, deriving the mathematical descriptions of these processes is challenging. In order to simplify these intricate physical processes, researchers have created physical models for DO and BOD. However, these physical models are still unable to predict DO and BOD with any degree of accuracy. Stochastic prediction models were developed because BOD and DO in rivers and streams change with time and show stochastic behavior. Regression models are most commonly used to estimate the stochastic behavior of BOD and DO. However, it is a difficult challenge to effectively replicate BOD and DO using typical regression models due to their very unpredictable behavior. When assessing water quality, prediction models are expected to possess a high degree of precognitive ability. Therefore, relying just on a basic statistical regression-based model to assess the quality of river water is not optimal.

Advanced artificial intelligence models comprise the current generation of computer-aided models. Artificial Intelligence is a dependable and extremely effective method for simulating groundwater and surface water quality. However, AI models showed robust and trustworthy modeling methods for a range of environmental, hydrological, and climatological applications. The primary advantage of AI models over traditional statistical methods, which are predicated on the idea of a linear link, is their ability to handle extremely complex nonlinear inter-factor correlations. The majority of studies have presented artificial intelligence (AI) models in a range of prediction model formats, including genetic programming, support vector machines, adaptive neuro-inference system models, and artificial neural networks (ANN).

This study aims to analyze water pollution in Texas City by leveraging advanced AI methodologies, specifically developing the Efficient AI-based Water Quality Prediction and Classification (EAI-WQP) model for accurate forecasting of key water quality parameters, implementing real-time big data processing using Apache Spark, optimizing model performance through Firefly Algorithm (FA), and classifying water quality using an Adaptive Neuro-Fuzzy Inference System (ANFIS); additionally, the research conducts comparative evaluations against traditional machine learning models such as GRU-ARIMA, SVM, and Random Forest to establish the EAI-WQP model's superiority in terms of accuracy, precision, F1 score, and recall, ensuring effective pollution monitoring and compliance with U.S. government standards.

In recent years, the integration of advanced technologies such as machine learning, big data, and artificial intelligence has significantly enhanced water quality monitoring and management. This literature survey explores cutting-edge research and developments in these areas, highlighting key advancements, methodologies, and their impact on improving water quality prediction and treatment.

Nair and Vijaya [1] examined river water quality prediction methods with an emphasis on machine learning and big data applications. It looks at several approaches, their efficacy, and how they might improve the forecasting of water quality. The paper emphasizes how crucial it is to combine big data and machine learning to enhance environmental management techniques and prediction accuracy. Fernando, et al. [2] investigated the use of artificial intelligence in wastewater treatment. It goes over the advantages and difficulties of using different AI algorithms to optimize treatment processes. The article demonstrates how increasing efficiency, sustainability, and system automation with AI can improve wastewater management. Arridha, et al. [3] offered an expansion of classification methods for real-time environmental monitoring utilizing big data analytics and the Internet of Things. It talks about a system that combines big data and IoT sensors to improve environmental monitoring and analysis, giving real-time insights and enhancing smart environment management.

Ghernaout, et al. [4] demonstrated how big data technologies are enabling enhanced management methods, increasing efficiency, and changing the way that water treatment processes are carried out. The potential of big data to transform water treatment techniques is discussed in the study along with developing trends. Kimothi, et al. [5] presented a framework for using machine learning, big data, and the Internet of Things to analyze indications of water quality. It talks about how various technologies work together to better monitor and forecast water quality, enhance real-time analysis, and advance water management techniques. Fu, et al. [6] used big data analysis and a water quality identification index to investigate long-term trends in surface water quality in China's main basins. It demonstrates how big data may be used to monitor and evaluate changes in water quality over time, offering insightful information about the condition of water bodies and the efficacy of management techniques. Sharma and Sharma [7] examined the application of IoT sensors and big data for real-time monitoring of water's physicochemical characteristics. The creation and use of smart sensors for ongoing surveillance is covered, as well as how big data analytics improve the management and interpretation of water quality data.

Budiarti, et al. [8] classified surface water using Support Vector Machines (SVM) in conjunction with big data technology. It illustrates how SVM may be used to categorize and evaluate surface water data using a case study from Surabaya, enhancing the assessment and management of water quality with sophisticated big data techniques. Rajaee, et al. [9] covered both single and hybrid approaches to AI-based models for river water quality prediction. It evaluates several AI approaches and how well they predict water quality, giving a thorough rundown of the most recent developments and cutting-edge practices in this area. Said [10] examined methods using artificial intelligence to forecast the quality of river water. In an effort to offer a thorough grasp of how AI can be applied in water quality prediction, it compiles different AI techniques, their applications, and their efficacy in forecasting water quality metrics. Madni, et al. [11] investigated the use of explainable AI and H2O AutoML for water-quality prediction. In order to improve prediction accuracy and model interpretability, it proposes a framework that blends explainable AI with automated machine learning. It focuses on how these methods improve the management of water quality.

Pandey and Verma [12] analyzed various AI approaches, including machine learning and deep learning, for the investigation and prediction of water quality. It outlines developments, uses, and potential paths for AI to enhance the evaluation and management of water quality. Abba, et al. [13] covered the use of ensemble machine learning and data intelligence models to forecast water quality indices. The article emphasizes the efficacy of integrating several machine learning methodologies to enhance prediction precision and presents case studies that showcase the potential of these models for water quality forecasting. Nallakaruppan, et al. [14] looked into employing explainable AI models to predict water quality with accuracy. It highlights how crucial the interpretability and transparency of the model are to guarantee reliable predictions. The article discusses several explainable AI approaches and how to use them to improve analyses and predictions of water quality.

Elkiran, et al. [15] explored the use of ensemble AI techniques for multi-step forward modeling of river water quality parameters. It highlights their potential to enhance long-term water quality forecasting and management by presenting approaches for projecting future water quality based on historical data and ensemble techniques. Ismail, et al. [16] compared the relationship between artificial intelligence and water treatment. It offers a comprehensive examination of AI applications in water treatment procedures, outlining advantages, difficulties, and areas in need of further research. The review seeks to provide a thorough grasp of the ways in which artificial intelligence might improve water treatment and management techniques.

Water contamination poses significant environmental and public health challenges, especially in regions experiencing rapid industrialization. In Texas City, USA, rivers such as the Red River, Rio Grande, and Trinity River have become critical indicators of water quality, prompting the need for innovative monitoring and predictive solutions. Advances in big data analytics and artificial intelligence offer promising avenues for real-time water quality management, ensuring compliance with stringent U.S. government standards for safe drinking water. This study builds on these technological strides by leveraging sophisticated data collection and analysis techniques to monitor essential water quality parameters, setting the stage for the development of a robust predictive model.

2. Experimental

2.1. Methods

2.1.1. Dataset

The investigation's data collection was based on information obtained from selective rivers such as the Red River, Rio Grande, and Trinity River in Texas City, United States, spanning from March 2023 to March 2024 to analyze water pollution. The dataset includes seven key parameters essential for water quality analysis: conductivity, pH, turbidity, dissolved oxygen (DO), total and faecal coliform, chemical oxygen demand (COD), and nitrate. Figure 1 shows the visual representation of the data collection areas. These parameters were monitored and analyzed to assess the water's condition and identify potential pollution sources. The U.S. government compiled all relevant data to ensure that drinking water met cleanliness standards and complied with safety regulations.

Figure 1.
Visual representation of (a) red river (b) rio grande river (c) trinity river.

2.2. Data Processing Tool

2.2.1. Apache Spark

Apache Spark is essential to the Data Processing stage because it manages and analyzes massive volumes of environmental sensor data in an effective manner. By utilizing its distributed computing capabilities, which enable the parallel execution of data processing tasks across several cluster nodes, it enables real-time processing. The capacity to manage the constant stream of data from environmental sensors is essential for ensuring that the data is processed accurately and quickly. Effective pollution control and intervention depend on prompt analysis and decision-making, which Apache Spark's real-time processing functionality facilitates. The approach makes sure that data handling and analytical processes are quick and responsive by leveraging Spark's optimized execution engine and in-memory processing. This allows for the latency-free development of actionable insights and forecasts.

2.3. Proposed Model

Advanced artificial intelligence techniques are utilized by efficient AI-based Water Quality Prediction and Classification (EAI-WQP) systems to precisely forecast and categorize water quality indicators. These systems are made to manage massive amounts of data from a variety of sources, including historical datasets, satellite photography, and sensors positioned in bodies of water. The ability to process and analyze complicated datasets is improved by the incorporation of AI models like machine learning (ML) and deep learning (DL), which makes it possible to find patterns and anomalies that older methods would have overlooked. Data pre-processing to eliminate noise, feature selection to find pertinent variables, and the use of predictive algorithms to anticipate water quality indicators like pH, turbidity, dissolved oxygen, and nutrient concentrations are important parts of EAI-WQP. The efficiency of the EAI-WQP model lies in its ability to handle and process large-scale, real-time data using Apache Spark, ensuring rapid and scalable analysis. Its parameter optimization using the Firefly Algorithm (FA) enhances predictive accuracy with minimal computational overhead. The integration of the ANFIS classifier ensures precise categorization of water quality while leveraging fewer resources compared to traditional methods.

EAI-WQP systems are very useful for managing and monitoring water resources in real time. These systems can offer early warnings of possible contamination events by using AI-driven models, which enables preventative actions to reduce pollution. They also make it easier to classify water quality into groups that can be used to inform public health alerts and regulatory compliance. The flexibility of AI models to various water bodies and environmental circumstances renders EAI-WQP a useful instrument for a range of ecosystems. These systems should get increasingly accurate and efficient as AI technology develops, greatly aiding in conservation and sustainable water management initiatives. Many factors that significantly affect the overall quality of the water are taken into account when calculating the Water Quality Index (WQI). During this investigation, a proposed model is assessed utilizing seven crucial water quality characteristics against a previously published dataset. To compute the WQI, the subsequent formula was employed in equation 1:

$$WI = \frac{\sum_{j=1}^{n} \times w_i}{\sum_{j=1}^{n} w_i} \quad (1)$$

On the quality rating scale I created for each parameter using the equation, the letter N stands for the number of elements taken into account while computing the WQI. The letter qi represents the quality rating for each parameter on this scale using the equation 2.

$$q_i = 100 \text{ X} \left(\frac{V_i - V_{ideal}}{S_i - V_{ideal}} \right) (2)$$

Table 1.
Permissible limits of the parameters used in calculating WQI.

S. No	Water Quality Parameter	Permissible limits
1	Ph	8.5
2	Turbidity	4 NTU
3	Dissolved Oxygen (DO)	12 mg/L
4	Conductivity	1000 S/m
5	Total Coliform	100 mL
6	Faecal Coliform	100 mL
7	Chemical Oxygen Demand (COD)	40 mg/L
8	Nitrate	9 mg/L

The measured metric i value, represented by the symbol Vi, was discovered in the tested water samples. The ideal value of the parameter Ideal is revealed by clean water, while Table 1 presents the recommended standard value of the parameter Si. The ideal value is called Ideal, and the suggested standard value is Si. Each water sample that was examined for analysis had a measured value for parameter i, which is called vi. The parameter i should be assigned the value V when the water is transparent. Figure 2 shows the architecture of the proposed methodology.

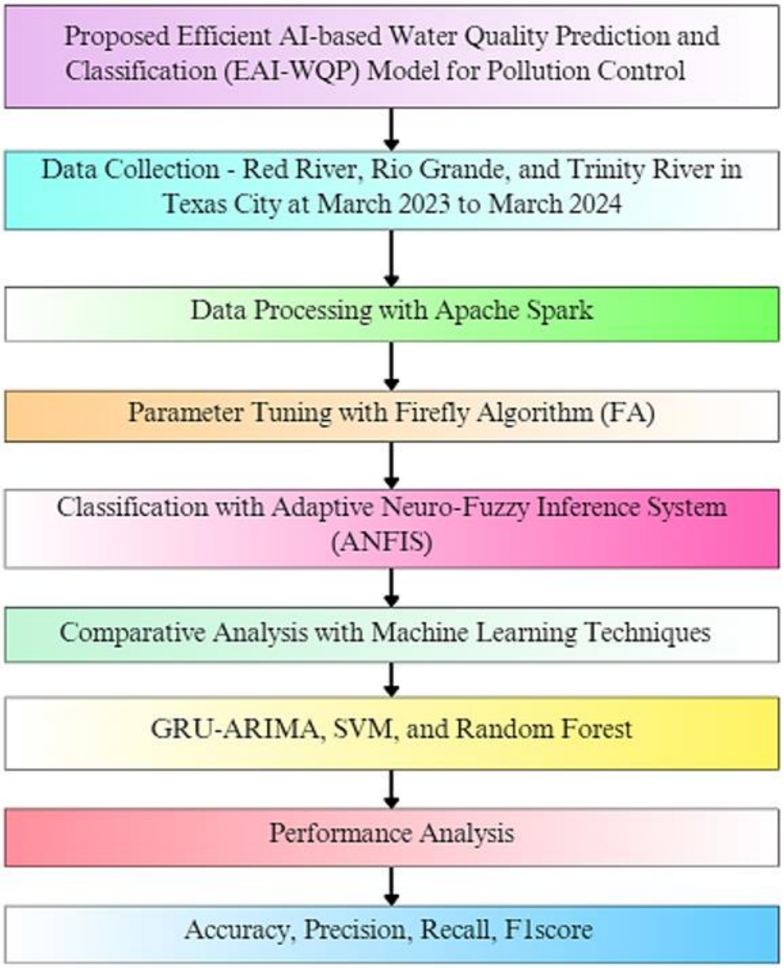
Proposed Efficient AI-based Water Quality Prediction and Classification (EAI-WQP) Model for Pollution Control
Data Collection - Red River, Rio Grande, and Trinity River in Texas City at March 2023 to March 2024
Data Processing with Apache Spark
Parameter Tuning with Firefly Algorithm (FA)
Classification with Adaptive Neuro-Fuzzy Inference System (ANFIS)
Comparative Analysis with Machine Learning Techniques
GRU-ARIMA, SVM, and Random Forest
Performance Analysis
Accuracy, Precision, Recall, F1score

Figure 2.
Proposed Methodology.

2.4. ML model

2.4.1. GRU-ARIMA

During the Model Development stage, a potent hybrid technique for predicting the levels of water pollution is created by combining the Autoregressive Integrated Moving Average (ARIMA) model with the Gated Recurrent Unit (GRU). To improve prediction accuracy and reliability, this hybrid model incorporates the best features of both conventional time series forecasting techniques and recurrent neural networks. Recurrent neural networks (RNNs), such as the GRU, are especially good at identifying patterns and temporal connections in sequential data. Its architecture, which is determined by the update equations below, enables it to efficiently manage long-term dependencies from equation 3 to 6:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (3)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (4)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \ (5)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \ (6)$$

Where, the update gate in this case is denoted by z_t, the reset gate by r_t, the candidate activation by h_t, and the concealed state by h_t. These formulas enable the GRU model to recognize intricate temporal patterns and learn from historical pollution data in an efficient manner. In contrast, the ARIMA model is a time series forecasting technique that has been around for a while and is well-known for its ability to predict univariate data with autocorrelated patterns. It can be express the ARIMA model as follows in equation 7:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \ (7)$$

Where Y_t is the observed value, ϕ denotes the autoregressive parameters, θ represents the moving average parameters, and ϵt is the white noise error term. This equation helps in capturing the autocorrelation within the pollution data, providing a robust baseline for predictions.

While the ARIMA component takes into consideration the data's seasonality and linear trends, the GRU component captures intricate, non-linear temporal connections. The hybrid model is expressed as follows in equation 8:

$$\hat{Y}_t = f\big(\text{GRU } (X_t)\big) + \text{ARIMA}(Y_{t-1}, \epsilon_{t-1}) \qquad (8)$$

where $\hat{Y}_t$ represents the predicted pollution levels, $f\big(\text{GRU } (X_t)\big)$ denotes the output from the GRU model based on input features X_t, and $\text{ARIMA } (Y_{t-1}, \epsilon_{t-1})$ represents the ARIMA component's

2.4.2. Parameter Tuning

The GRU-ARIMA model is optimized using the Firefly Algorithm (FA) during the Parameter Tuning phase to improve the model's accuracy in forecasting water pollution levels. The Firefly Algorithm is a metaheuristic optimization method for handling challenging optimization problems. It was inspired by the flashing characteristic of fireflies. The idea behind it is that brighter firefly, which stand for better solutions. A population of fireflies, each of which represents a potential solution to the optimization issue, is initialized as part of the FA algorithm's operation. The objective function, in this case the GRU-ARIMA model's forecasting accuracy, determines each firefly's brightness. The i-firefly's brightness, Bi, can be stated as follows in equation 9:

$$B_i = -\text{MSE}_i \quad (9)$$

where MSEi stands for the model's Mean Squared Error and its parameters correspond to the firefly's parameters. Higher brightness is correlated with a lower mean square error (MSE). A firefly i's attractiveness (Aij) to a firefly j is defined as follows in equation 10:

$$A_{ij} = \beta_0 \exp\left(-\gamma r_{ij}^2\right) (10)$$

where β_0 is the attractiveness at rij=0, γ is the light absorption coefficient, and r_{ij} is the distance between fireflies i and j. The movement of a firefly iii towards a more attractive firefly j is given in equation 11:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \alpha \cdot \left(\mathbf{x}_j - \mathbf{x}_i\right) + \epsilon \ (11)$$

where $x_i(t+1)$ is the updated position of firefly i, α is a random step size, and ϵ is a random perturbation. This movement allows fireflies to search the parameter space more effectively, improving the model's accuracy.

The GRU-ARIMA model's hyperparameters, including the GRU's learning rate, the ARIMA model's orders, and any regularization parameters, are iteratively adjusted via the Firefly Algorithm. The model's parameters are fine-tuned to produce optimal forecasting performance through the FA, which minimizes the Mean Squared Error (MSE) or other pertinent objective functions. By determining the most efficient parameter settings, this optimization method improves the GRU-ARIMA hybrid model's accuracy, resulting in more accurate water pollution level predictions and better overall model performance.

2.5. Data classification

Adaptive Neuro-Fuzzy Inference System (ANFIS) is used in the Classification phase to divide water quality into groups of pollutants and non-pollutants. ANFIS is a potent hybrid system that effectively handles complicated and nonlinear environmental data by fusing the interpretability of fuzzy logic with the learning capabilities of neural networks. The system is set up as a five-layer neural network, with pH, turbidity, and chemical concentrations being sent to the input layer of the network. These inputs are transformed into fuzzy sets in the fuzzification layer by membership functions, which can manage the imprecision and uncertainty present in environmental data. Fuzzy if-then rules that link input conditions to output categories comprise the rule layer. These rules contain expert knowledge regarding the classification of water quality. ANFIS uses a backpropagation method in conjunction with a least squares estimation to improve these rules and the related parameters during the learning process. The goal is to minimize the classification error. The defuzzification layer provides an unambiguous classification of water quality as either pollutant or non-pollutant by converting the fuzzy outputs back into crisp values. By combining these elements, ANFIS is able to accurately classify and analyze complicated environmental data with ease, enabling efficient monitoring and management of water quality that is impacted by industrial activity.

3. Results and Discussion

3.1. Accuracy

The accuracy of water quality parameters assessed by different models shows a notable performance difference which is illustrated in Figure 3. For this experiment, samples were obtained from 3 rivers from Texas City in UN from March 2023 to March 2024. For pH measurement, the ML model achieved 88% accuracy, while the EAI-WQP system, using ANFIS+FIREFLY, reached 95%. Turbidity was accurately predicted at 85% by the ML model compared to 92% by the EAI-WQP system. Dissolved Oxygen (DO) measurements had an accuracy of 86% with the ML model, whereas the EAI-WQP system improved this to 94%. Conductivity accuracy was 87% with the ML model and 93% with the EAI-WQP system.

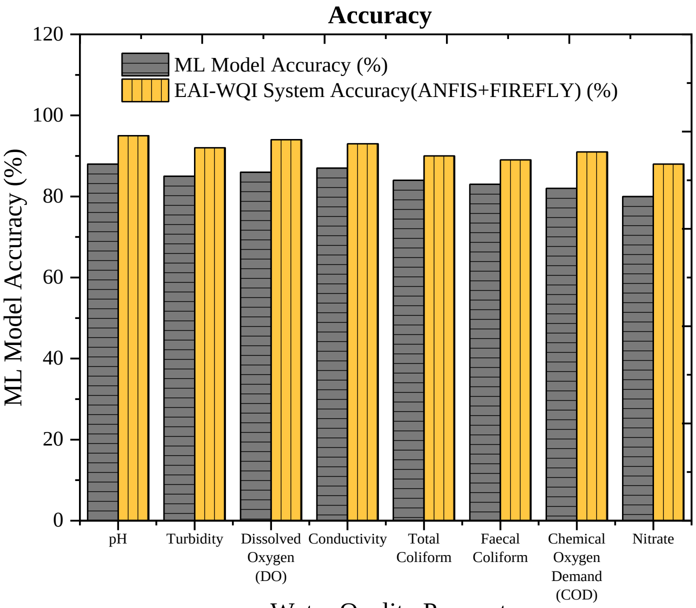

Figure 3.
Accuracy results of ML Model vs. EAI-WQP System.

For Total Coliform, the ML model's accuracy was 84%, while the EAI-WQP system achieved 90%. Faecal Coliform predictions were 83% accurate with the ML model and 89% with the EAI-WQP system. Chemical Oxygen Demand (COD) showed an accuracy of 82% with the ML model, and 91% with the EAI-WQP system. Lastly, Nitrate levels were predicted with 80% accuracy by the ML model and 88% by the EAI-WQP system. Overall, the EAI-WQP system consistently outperforms the ML model across all water quality parameters.

3.2. Precision

Figure 4 compares the precision of ML models and the EAI-WQP system, which uses ANFIS and Firefly Algorithm. The EAI-WQP system consistently outperformed the ML model across all parameters. For pH, the ML model achieved 85% precision, while EAI-WQP reached 93%. Turbidity precision was 82% for the ML model and 90% for EAI-WQP. Dissolved Oxygen (DO) had 84% precision with the ML model and 91% with EAI-WQP. Conductivity, Total Coliform, Faecal Coliform, Chemical Oxygen Demand (COD), and Nitrate also showed higher precision with EAI-WQP compared to the ML model.

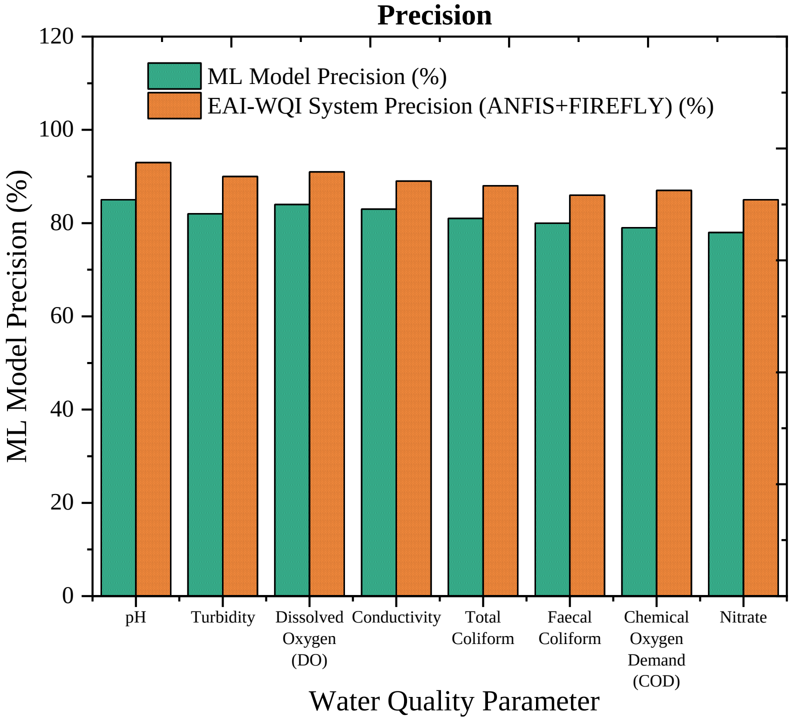

Figure 4.
Precision results of ML Model vs. EAI-WQP System.

3.3. Recall

The comparison between the machine learning (ML) model and the EAI-WQP system (which combines ANFIS and the Firefly algorithm) across various water quality parameters reveals a generally superior performance of the EAI-WQP system. Specifically, for pH, the EAI-WQP system achieved a recall of 92%, outperforming the ML model's 84%. Similarly, the EAI-WQP system showed higher recall rates for turbidity (91% vs. 83%), dissolved oxygen (93% vs. 85%), conductivity (90% vs. 82%), total coliform (89% vs. 80%), faecal coliform (87% vs. 78%), chemical oxygen demand (88% vs. 81%), and nitrate (86% vs. 79%). This suggests that the EAI-WQP system provides a more accurate assessment of water quality across these parameters compared to the ML model. Recall performance was seen in Figure 5.

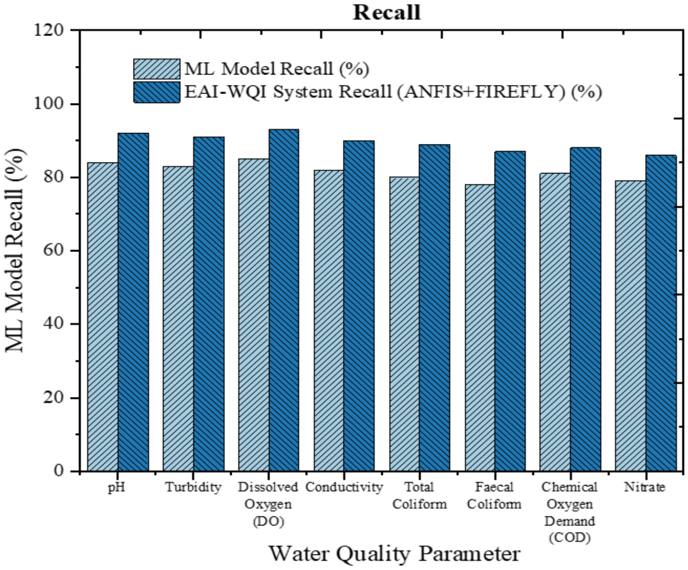

Figure 5.
Recall results of ML Model vs. EAI-WQP System.

3.4. F1 Score

Figure 6 compares F1 scores for various water quality parameters between ML models and the EAI-WQP system, which uses ANFIS and Firefly Algorithm. The EAI-WQP system outperforms the ML model in most cases: pH (92% vs. 86%), Turbidity (91% vs. 84%), and Dissolved Oxygen (92% vs. 85%). The ML model scores higher for Conductivity (83% vs. 80%), while the EAI-WQP system leads for Total Coliform (88% vs. 81%), Faecal Coliform (86% vs. 79%), COD (89% vs. 82%), and Nitrate (87% vs. 80%).

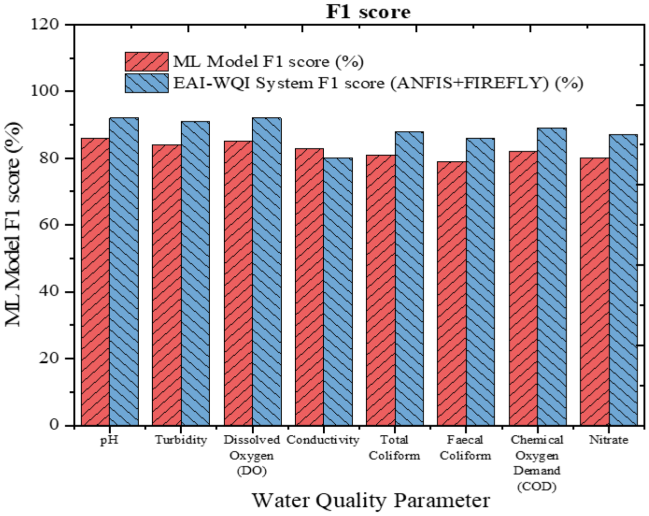

Figure 6.
F1 Score results of ML Model vs. EAI-WQP System.

3.5. Comparative Analysis of Proposed and Existing ML Model

Table 2 and Figure 7 displayed the performance metrics of various models that reveal distinct levels of effectiveness. The Proposed EAI-WQP model, utilizing Firefly and ANFIS, demonstrated the highest performance with an accuracy of 94.5%, precision of 92.3%, recall of 93.7%, and an F1-Score of 93.0%. The GRU-ARIMA model followed with an accuracy of 88.7%, precision of 86.5%, recall of 87.0%, and an F1-Score of 86.7%. The SVM model exhibited an accuracy of 85.2%, precision of 83.4%, recall of 84.1%, and an F1-Score of 83.7%. Lastly, the Random Forest model recorded an accuracy of 87.9%, precision of 85.7%, recall of 86.3%, and an F1-Score of 86.0%. Overall, the Proposed EAI-WQP model outperformed the other models in all evaluated metrics.

GRU-ARIMA, SVM, and Random Forest were chosen as comparison models because they represent different approaches to prediction, offering a diverse evaluation of the proposed model's performance. GRU-ARIMA combines the strengths of deep learning (GRU) and traditional time series forecasting (ARIMA), making it suitable for predicting water quality trends. SVM is known for its ability to handle high-dimensional data and perform well in classification tasks, providing a solid baseline for comparison in terms of classification accuracy. Random Forest, an ensemble learning method, is widely used for its robustness and ability to handle large datasets with complex patterns. By comparing EAI-WQP against these established models, the study evaluates its relative performance and demonstrates its superiority in terms of accuracy, precision, and recall.

Table 2.
Comparative Performance Metrics.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed EAI-WQP (firefly + ANFIS)	94.5	92.3	93.7	93.0
GRU-ARIMA	88.7	86.5	87.0	86.7
SVM	85.2	83.4	84.1	83.7
Random Forest	87.9	85.7	86.3	86.0

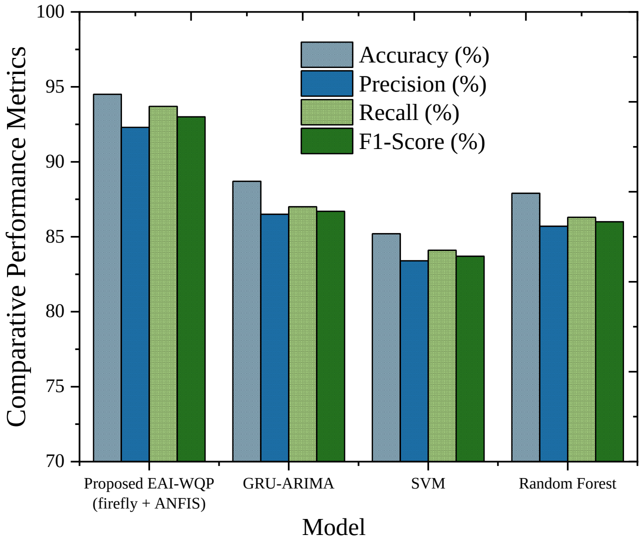

Figure 7.
Comparative results of ML models and proposed technique.

4. Limitations

Despite the promising results, the study has several limitations. It relies on data from only three rivers in Texas City, which may not fully represent diverse environmental conditions and pollution

DOI: 10.55214/25768484.v9i7.8654

sources across different regions. The complexity of integrating the Firefly Algorithm and ANFIS may also require substantial computational resources, limiting real-time application in constrained systems. Furthermore, the study did not compare the model with more advanced deep learning techniques like LSTMs or CNNs, leaving room for potential improvements. Future research should expand the dataset to include a broader range of water bodies, incorporate real-time sensor data and satellite imagery, and explore advanced deep learning models for improved prediction accuracy. Optimizing the Firefly Algorithm for efficiency and combining machine learning techniques could enhance performance, while including socio-economic and climate data could provide a more comprehensive approach to water quality prediction in urban settings.

5. Conclusion

The study analyzed water quality parameters from the Red River, Rio Grande, and Trinity River in Texas City, USA, from March 2023 to March 2024. The proposed EAI-WQP model outperforms benchmarks due to its integration of real-time big data processing with Apache Spark and optimization via the Firefly Algorithm (FA). Its Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier enhances classification accuracy by combining fuzzy logic with neural networks. Comparative evaluations show superior performance in accuracy, precision, F1 score, and recall over GRU-ARIMA, SVM, and Random Forest models.

It was evaluated against GRU-ARIMA, SVM, and Random Forest models. The following results were obtained from the below research:

1. The ML model achieved 88% accuracy for pH, 85% for turbidity, 86% for Dissolved Oxygen (DO), 87% for conductivity, 84% for Total Coliform, 83% for Faecal Coliform, 82% for Chemical Oxygen Demand (COD), and 80% for nitrate. In contrast, the EAI-WQP system reached 95% accuracy for pH, 92% for turbidity, 94% for DO, 93% for conductivity, 90% for Total Coliform, 89% for Faecal Coliform, 91% for COD, and 88% for nitrate.
2. The ML model's precision for pH was 85%, for turbidity was 82%, and for DO was 84%. The EAI-WQP system outperformed with 93% precision for pH, 90% for turbidity, and 91% for DO. Precision was consistently higher for all parameters with the EAI-WQP system compared to the ML model.
3. The ML model achieved an F1 score of 86% for pH, 84% for turbidity, and 85% for DO. The EAI-WQP system showed higher F1 scores of 92% for pH, 91% for turbidity, and 92% for DO. The EAI-WQP system had superior F1 scores across all parameters.
4. The ML model had a recall rate of 84% for pH, 83% for turbidity, and 85% for DO. The EAI-WQP system demonstrated higher recall rates of 92% for pH, 91% for turbidity, and 93% for DO. The EAI-WQP system consistently outperformed the ML model in recall across all parameters.
5. The EAI-WQP system consistently outperformed the ML model in accuracy, precision, F1 score, and recall across all water quality parameters. The EAI-WQP system showed significant improvements over the ML model, achieving higher performance metrics in every evaluated category.

Despite the promising results of the EAI-WQP model, several limitations must be acknowledged. First, the study relies on data collected from only three rivers in Texas City, which may limit the model's generalizability to other geographic regions with different hydrological and environmental conditions. Additionally, while the Firefly Algorithm (FA) enhances parameter optimization, its computational complexity could pose challenges for large-scale implementations. The reliance on historical data may also introduce biases, as sudden pollution events or emerging contaminants not represented in the dataset could affect real-time predictive accuracy. Future research should focus on expanding the dataset to include diverse water bodies across multiple climatic regions, integrating additional water quality indicators such as heavy metals and microplastics, and refining the AI model with hybrid optimization techniques for improved scalability. Moreover, incorporating Internet of Things (IoT) sensors for continuous real-time monitoring and integrating the model with decisionsupport systems for policymakers could enhance its practical applicability in water resource management.

Abbreviation:

EAI-WQP	-	Efficient AI-Based Water Quality Prediction and Classification
GRU-ARIMA	-	Gated Recurrent Unit - Autoregressive Integrated Moving Average
ANFIS	-	Adaptive Neuro-Fuzzy Inference System
FA	-	Firefly Algorithm
ANN	-	Artificial Neural Networks
DO	-	Dissolved Oxygen
ML	-	Machine Learning
DL	-	Deep Learning
WQI	-	Water Quality Index
GRU	-	Gated Recurrent Unit
ARIMA	-	Autoregressive Integrated Moving Average
RNN	-	Recurrent Neural Network
SVM	-	Support Vector Machine
MSE	-	Mean Squared Error

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Acknowledgment:

I would like to express my sincere gratitude to my co-authors of this research. Special thanks to the institution that made this work possible. I also appreciate the constructive feedback and insightful comments from my co-authors, which greatly enhanced the quality of this manuscript.

Copyright:

References

[1] J. P. Nair and M. Vijaya, "Predictive models for river water quality using machine learning and big data techniques-a Survey," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021: IEEE, pp. 1747-1753.

[2] W. A. M. Fernando, S. N. Khadaroo, and P. E. Poh, "Artificial intelligence in wastewater treatment systems in the era of industry 4.0: A holistic review," *Artificial Intelligence and Environmental Sustainability: Challenges and Solutions in the Era of Industry*, vol. 4, no. 4, pp. 45-85, 2022.

[3] R. Arridha, S. Sukaridhoto, D. Pramadihanto, and N. Funabiki, "Classification extension based on IoT-big data analytic for smart environment monitoring and analytic in real-time system," *International Journal of Space-Based and Situated Computing*, vol. 7, no. 2, pp. 82-93, 2017. https://doi.org/10.1504/ijssc.2017.086821

[4] D. Ghernaout, M. Aichouni, and A. Alghamdi, "Applying big data in water treatment industry: A new era of advance," *International Journal of Advanced and Applied Sciences*, vol. 5, no. 3, pp. 89–97, 2018. https://doi.org/10.21833/ijaas.2018.03.013

[5] S. Kimothi *et al.*, "Big data analysis framework for water quality indicators with assimilation of IoT and ML," *Electronics*, vol. 11, no. 13, p. 1927, 2022.

[6] X. Fu, R. Wu, H. Qi, and H. Yin, "Long-term trends in surface water quality of China's seven major basins based on water quality identification index and big data analysis," *Environmental Impact Assessment Review*, vol. 100, p. 107090, 2023. https://doi.org/10.1016/j.eiar.2023.107090

[7] N. Sharma and R. Sharma, "Real-time monitoring of physicochemical parameters in water using big data and smart IoT sensors," *Environment, Development and Sustainability*, vol. 4, no. 10, pp. 1–48, 2022.

[8] R. P. N. Budiarti, S. Sukaridhoto, M. Hariadi, and M. H. Purnomo, "Big data technologies using SVM (case study: Surface water classification on regional water utility company in Surabaya)," in *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, 2019: IEEE, pp. 94-101.

[9] T. Rajaee, S. Khani, and M. Ravansalar, "Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 200, p. 103978, 2020.

[10] M. I. M. Said, "Artificial intelligence approach to predicting river water quality: A review," *Journal of Environmental Treatment Techniques*, vol. 8, no. 3, pp. 1093-1100, 2020.

[11] H. A. Madni *et al.*, "Water-quality prediction based on H2O AutoML and explainable AI techniques," *Water*, vol. 15, no. 3, p. 475, 2023. https://doi.org/10.3390/w15030475

[12] J. Pandey and S. Verma, "Water quality analysis and prediction techniques using artificial intelligence," in *ICT with Intelligent Applications: Proceedings of ICTIS 2021, Volume 1*: Springer, 2021, pp. 279-290.

[13] S. I. Abba *et al.*, "Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index," *Environmental Science and Pollution Research*, vol. 27, no. 33, pp. 41524-41539, 2020. https://doi.org/10.1007/s11356-020-09689-x

[14] M. K. Nallakaruppan, E. Gangadevi, M. L. Shri, B. Balusamy, S. Bhattacharya, and S. Selvarajan, "Reliable water quality prediction and parametric analysis using explainable AI models," *Scientific Reports*, vol. 14, no. 1, p. 7520, 2024. https://doi.org/10.1038/s41598-024-56775-y

[15] G. Elkiran, V. Nourani, and S. I. Abba, "Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach," *Journal of Hydrology*, vol. 577, p. 123962, 2019. https://doi.org/10.1016/j.jhydrol.2019.123962

[16] W. Ismail, N. Niknejad, M. Bahari, R. Hendradi, N. J. M. Zaizi, and M. Z. Zulkifli, "Water treatment and artificial intelligence techniques: A systematic literature review research," *Environmental Science and Pollution Research*, vol. 30, no. 10, pp. 1–19, 2021.

DOI: 10.55214/25768484.v9i7.8654