An Artificial Neural Network Model for Water Quality Prediction in the Amoju Hydrographic Subbasin, Cajamarca-Peru

Alex Alfredo Huaman Llanos[1], Jeimis Royler Yalta Meza[2], Danicza Violeta Sanchez Cordova[3], Juan Carlos Chasquero Martinez[4], Lenin Quiñones Huatangari[5], Dulcet Lorena Quinto Sanchez[6], Roxana Rojas Segura[7], Alfredo Lazaro Ludeña Gutierrez[8]

Informatic and Language Center, National University of Jaen, Jaen, Peru[1]
Direction of Production Center of Goods and Services, National University of Jaen, Jaen, Peru[2]
Soil and Water Analysis Center, National University of Jaen, Jaen, Peru[3, 6]
Geospatial Analysis and Computing Laboratory, National University of Jaen, Jaen, Peru[4]
Professional School of Data Science Engineering and Artificial Intelligence,
Toribio Rodriguez de Mendoza National University, Chachapoyas, Peru[5]
Biotechnology-Genetics and Molecular Biology Laboratory, National University of Jaen, Jaen, Peru[7]
Faculty of Engineering, National University of Piura, Piura, Peru[8]

***Abstract*—Water quality is crucial for sustaining life, and accurate prediction models are essential for effective management. This study introduces an Artificial Neural Network (ANN) model designed to predict the Water Quality Index (WQI) in the Amoju Hydrographic Subbasin, Cajamarca-Peru. The model was developed using key water quality parameters, including electrical conductivity (EC), total dissolved solids (TDS), calcium carbonate ($CaCO_3$), and phosphate (PO_4^{3-}), identified through Pearson correlation analysis. Data from water samples collected over six months were used to train and validate the model. Results revealed that the ANN model achieved high predictive accuracy, with a significant correlation between WQI and the aforementioned parameters. The model's performance outstrips traditional methods demonstrating its capability to effectively capture complex interdependencies among water quality indicators. This research emphasizes the potential of AI-driven approaches for enhancing predictive accuracy in environmental monitoring. Future studies should consider incorporating additional variables, such as heavy metals and microbial indicators, and consider the application of real-time AI-driven monitoring systems to further refine water quality management strategies. The ANN model presented here offers a promising tool for decision-makers, providing a reliable method for predicting water quality in similar hydrographic basins and contributing to the broader field of AI in environmental science.**

Keywords*—Artificial neural networks; hydrographic subbasin; machine learning models; water quality index; water resource management

I. INTRODUCTION

Water is a critical resource upon which all life on Earth depends, with its quality playing a pivotal role in human health and aquatic ecosystems. The contamination of water sources poses significant threats to public health and the environment, underscoring the urgency of effective water quality monitoring and management. Numerous studies have highlighted the importance of assessing and predicting water quality to safeguard the sustainability and safety of water resources [1], [2]. Water quality forecasting is an indispensable method for effective water resource planning, regulation, and monitoring, and it is a crucial component of research focused on water ecological protection [3], [4].

The global issue of water pollution, aggravated by industrialization and urbanization, has driven the development of advanced methodologies for monitoring and predicting water quality. Traditional approaches, such as manual sampling followed by laboratory analysis, are often labor-intensive, time-consuming, and costly, thereby limiting their efficiency and scalability. These challenges have catalyzed the integration of artificial intelligence (AI) and machine learning (ML) techniques into water quality assessment, offering more efficient and cost-effective solutions for real-time water quality prediction [5], [6]. Among these techniques, machine learning models, particularly artificial neural networks (ANNs), have gained widespread adoption due to their capability to handle complex, and nonlinear relationships within environmental data [7], [8].

Recent advancements in technology, including remote sensing (RS), the Internet of Things (IoT), and big data analytics, have further enhanced water quality monitoring by enabling the collection and processing of vast amounts of data form diverse source. The synergy between these technologies and AI has facilitated the development of more accurate and reliable predictive models, capable of providing comprehensive assessments of water quality status [9], [10]. Specifically, metrics such as the Water Quality Index (WQI) and Water Quality Classification (WQC) are commonly employed to aggregate multiple water parameters into a single, interpretable value, providing a holistic overview of water quality [11].

This study is focused on developing an Artificial Neural Network (ANN) model to predict water quality in the Amoju Hydrographic Subbasin located in Cajamarca Peru. By leveraging various water quality indicators, the model aims to forecast the WQI and classify the water quality status, thereby contributing valuable insights for water resource management and pollution control. Through the application of advanced AI techniques, this research addresses the pressing need for efficient water quality prediction methods that ensure the provision of safe and clean water for diverse uses, while also mitigating the adverse effects of water contamination on public health and the environment [12], [13].

The structure of the paper is organized as follows: Section 2 reviews the relevant literature on water quality prediction using various classifiers. Section 3 details the materials and methodologies employed, including data preparation, pre-processing, splitting, distribution, feature correlation, and WQI computation. The experimental setup and the result analysis are discussed in Section 4. Finally, the paper is concluded with limitations and future scope in Section 5.

II. Literature Review

The literature on water quality prediction has undergone a considerable transformation, particularly with the increasing adoption of Artificial Neural Networks (ANNs) in environmental modeling. This paradigm shift is evident in the application of ANNs within hydrographic subbasins, such as the Amoju Subbasin in Cajamarca, Peru. The study titled "An Artificial Neural Network Model for Water Quality Prediction in the Amoju Hydrographic Subbasin, Cajamarca-Peru" contributes significantly to this evolving field by addressing the critical need for accurate predictive models that can inform environmental management and policy in the region. ANNs, inspired by the neural structures of the human brain, excel at capturing and modeling the intricate non-linear relationships prevalent in environmental datasets. These networks are particularly effective in dealing with the complex dynamics of hydrographic subbasins, such as the Amoju Subbasin, which holds significant ecological value and is vulnerable impacts of anthropogenic activities.

Recent studies further underscore the potential of artificial intelligence (AI) in enhancing water quality prediction and monitoring. For instance, [14] and [15] developed AI models focusing on water quality index (WQI) prediction and water quality classification (WQC). The study in [14] utilized adaptive neuro-fuzzy inference system (ANFIS) algorithms for WQI prediction and feed-forward neural network (FFNN) for WQC, achieving high predictive accuracy. The study in [15] also demonstrated the efficacy of AI in water quality monitoring, affirming the robustness of these models in managing complex environmental data. The research in [16] reviewed the integration of AI and the Internet of Things (IoT) in water quality prediction, emphasizing AI's role in analyzing intricate systems and leveraging historical data to enhance prediction accuracy. Similarly, [17] explored several AI techniques, including multilayer perceptron neural networks (MLP-ANN), ensemble methods, gaussian process regression, support vector machine (SVM), and decision tree, all of which contributed to a comprehensive evaluation of water quality parameters.

The impact of water quality on public health further emphasizes the necessity of accurate WQI modeling, a task that presents significant challenges within the water sector. The study in [18] introduced an innovative application of the ensemble Kalman filter integrated with ANNs to predict WQI using physicochemical parameters, showcasing the model's capability to handle noise in environmental data. Other researchers have explored different machine learning techniques for WQI prediction. The research in [19] employed k-nearest neighbors, boosting decision trees, SVMs, and multilayer perceptron ANNs in their models, while [20] compared deep learning-based models with other machine learning models such as random forests (RF) and extreme gradient boosting (XGBoost) for predicting groundwater quality. Their comparative study highlighted the superior performance of deep learning models in certain contexts.

The prediction of dissolved oxygen concentration, a key indicator of river water quality, has also been enhanced through AI. The study in [21] utilized a deep learning approach, applying recurrent neural networks (RNNs) to predict this parameter with high precision. Moreover, [22] proposed a hybrid model combining ANNs, discrete wavelet transforms (DWT), and long short-term memory (LSTM) networks, further advancing the state-of-the-art in water quality prediction.

In addition, [23] implemented a comprehensive architecture integrating machine learning models (RF, DT, LR, SVM, AdaBoost) with deep learning models (CNN, LSTM, GRU) for predicting both water quality and water consumption. This study underscores the potential of hybrid models in addressing multi-faceted environmental issues. The research in [24] further advanced this approach by integrating deep learning models with feature extraction technique, such as principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA), to enhance water quality classification accuracy.

In study [25], focused on the prediction of water quality using machine multiple learning techniques, including regression and classification models like SVMs, multiple linear regression (MLR), and Bayesian tree model (BTM). Their comprehensive five-step methodology, which included data pre-processing, feature correlation analysis, and model feature importance, resulted in a maximum prediction accuracy of 99.83% with the MLR classifier. The study in [26] compared decision tree algorithms (DT) and Naïve Bayes classifiers for water quality prediction, with DT emerging as the most accurate model, achieving an accuracy value of 97.23%.

Additionally, [27] proposed a machine learning-based system for predicting the WQI in the Illizi region by employing eight artificial intelligence algorithms. The results indicated that the Multivariate Linear Regression (MLR) model exhibited the highest accuracy among the models considered. In contrast, [28] explored 15 supervised Machine Learning (ML) algorithms to estimate the WQI, identifying gradient boosting and polynomial regression as the most efficient methods for WQI prediction. Further advancements in the field

include the work of [29], who developed a prediction method for WQI using feedforward artificial neural networks (ANNs) with 25 water quality parameters inputs. By integrating backward elimination and forward selective combination methods, the study achieved high R2 and minimal squared error (MSE). Similarly, [30] predicted the WQI using 16 water quality parameters and successfully applied ANN through a Bayesian regularization algorithm, demonstrating the robustness of this approach. Moreover, the study in [31] examined the comparative efficiency of multivariate linear regression (MLR) and ANN models for predicting water quality parameters, such as pH, temperature, total suspended solids (TSS), and total suspended matter (TSM), in estimating the chemical oxygen demand (COD) and biochemical oxygen demand (BOD). In another study, [32] designed a feed-forward, fully-connected, three-layer perceptron neural network model to predict the WQI using 23 parameters, reinforcing the trend towards increasingly complex ANN architectures. Also, [33] took a different approach by proposing a two-layered ensemble model that integrates five commonly used methods, including partial least square, random forest, and Bayesian networks, into an ML model for forecasting beach water quality. The model stacking approach yielded the best predictions, demonstrating the advantages of ensemble methods in enhancing model robustness and accuracy. The study in [34] developed an ML-based classification system for the Chao Phraya River's water quality, integrating attribute realization (AR) and support vector machine (SVM) algorithms. The results indicated that linear regression (LR) was the most suitable function for river water data classification, offering a different perspective on the adaptability of ML techniques across diverse water bodies. The research in [35] also employed an ANN approach for calculating and simulating the WQI of the Akaki River, utilizing a neural network model trained on 12 inputs and one output. The optimal model architecture was obtained with eight hidden layers, achieved an accuracy of 0.93. Besides, [36] formulated four distinct ML techniques, including Back Propagation Neural Network (BPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Multilinear Regression (MLR), and Support Vector Regressor (SVR) for forecasting the water quality index (WQI) across the Yamuna River. The study in [37] applied independent techniques like the M5P tree model, additive regression (AR), support vector machine (SVM), and random subspace (RSS) to predict WQI, identifying AR as the most optimal approach with favorable outcomes.

The literature also explores hybrid models. The study in [38] developed a hybrid ML method combining random trees and bagging, testing four standalone and 12 hybrid data-mining algorithms for WQI forecasting in a humid climate. The study concluded that hybrid models could significantly improve prediction accuracy. In a similar vein, [39] optimized the performance of an adaptive neuro-fuzzy inferences system (ANFIS) for water quality metrics prediction using Genetic Algorithm (GA), Differential Evolution (DE), and Ant Colony Optimization (ACOR), further demonstrating the value of optimization algorithms with ML models. Consequently [40] enhanced a hybrid artificial neural network (HANN) model with a genetic algorithm (GA) for predicting water output in drinking water treatment plants in China. The HANN model has shown better ability and consistency in forecasting the total water output. The prediction shows that the HANN model has improved its performance from 0.71 to 0.93 R2 by increasing the training data provided. Likewise, the study in [41] introduced an ensemble ML model, Extra Tree Regression (ETR), for predicting monthly WQI values in Hong Kong. Achieving a high prediction accuracy with R2 = 0.98 and RMSE = 2.99. The study in [42] utilized Principal Component Regression (PCR) and Gradient Boosting Classifier (GBC) to predict WQI, demonstrating 95% prediction accuracy for PCR method and 100% classification accuracy for GBC. [43] evaluated the performance of 12 ML models, including boosting-based, decision tree-based, and ANN-based algorithms, for estimating the WQI of the La Boung River in Vietnam, with extreme gradient boosting (XGBoost) emerging as the best performer, achieving an R2 of 0.989 and RMSE of 0.107. Finally, [44] used Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Adaptive Boosting (Ada-Boost) model for predicting WQC. In contrast, K-nearest neighbor (KNN), decision tree (DT), support vector regressor (SVR), and multi-layer perceptron (MLP) were used as regression models for predicting WQI. The results showed that GB model produced the best results for predicting WQC, with an accuracy of 99.5% value, and the MLP regressor model in predicting WQI, with an accuracy of 99.8% value.

These studies collectively highlight the potential of AI, particularly machine learning and neural networks, in advancing water quality management. However, there is a notable gap in evaluating water quality classification based on metrics such as accuracy, precision, or F1 score. Moreover, many of these approaches are limited by their focus on either WQI prediction or WQC, rather than integrating both aspects. Our proposed methodology addresses these aspects by employing a lightweight model that not only enhances prediction accuracy but also integrates water quality classification with water demand prediction, paving the way for a more comprehensive approach to water resource management. Table I provides a comparative summary of the research works discussed.

The literature further reveals the versatility of artificial neural networks (ANNs) in water quality prediction, demonstrating their adaptability to various geographical and hydrological contexts. The inclusion of diverse input parameters, including meteorological data to land use and physicochemical attributes-into ANN models has consistently improved the prediction accuracy, offering a more holistic understanding of the factors influencing water quality dynamics.

TABLE I. COMPARATIVE SUMMARY OF THE DISCUSSED RESEARCH WORKS

Reference	Year	Classifiers	Achieved Accuracy
[14]	2020	NARNET, LSTM, SVM, KNN, NB	SVM 97.01%
[15]	2021	KNN, FFNN, ANIFS	WQI ANFIS 96.17%
[24]	2021	Dimension reduction PCA, LDA, ICA, RNN, LSTM, SVM (variants)	LSTM RNN with LDA, LSTM, RNN 99.72%
[25]	2021	NN, RF, MLR, SVM, BT	99.83% MLR
[26]	2021	DT, NB (variants), K-fold cross validation	97.22% DT
[27]	2021	MLR, RF, RSS, AR, ANN, SVR, LWLR	With all parameters: MLR
[28]	2019	Multiple linear regression, polynomial regression, RF, GBC, SVM, ridge regression, lasso regression, elastic net regression, MLP, GNB, LR, SGD, KNN, DT, bagging classifier	Gradient Boosting and Polynomial Regression with MAE = 1.964 and 2.727, respectively
[34]	2021	Attribute-realization (AR) and Support Vector Machine (SVM)	AR-SVM with 0.86-0.95 accuracy respectively
[36]	2019	BPNN, ANFIS, SVR and MLR	DC values vary in the range of 0.9202 to 0.9957
[37]	2022	AR, M5P tree model, RSS and SVM	AR with R2 = 0.9993, MAE = 0.5243, RSME = 0.6356, %RAE = 3.8449 and %RRSE = 3.9925
[38]	2020	RF, M5P, RT and REPT	RT with R2 = 0.941, RMSE = 2.71, MAE = 1.87, NSE = 0.941, PBIAS = 0.500
[41]	2021	ETR, SVR and DTR	ETR model produced more accurate WQI predictions with R2 = 0.98 and RSME = 2.99 values
[42]	2022	PCR and GBC methods	PCR = 95% and GBC = 100%
[43]	2022	(Adaptive boosting, GBoost, HGBoost, LGBoost, XGBoost), (DT, ET, RF), (MLP, RBF, DFFNN, CNN)	XGBoost (R2 = 0.989 and RMSE = 0.107)
[44]	2023	K-nearest neighbor (KNN) regressor model, DT, SVR, MLP	MLP regressor model outperformed the best accuracy with R2 = 99.8%

III. MATERIALS AND METHODOLOGY

Fig. 1 displays the proposed methodology for completing the research.

A. Study Area

The study was conducted in the Amoju River Subbasin, which is located within the Alto Marañon III Inter-basin in northern Peru. The subbasin covers an area of 354 km2, as depicted in Fig. 2. The Amoju River itself extends approximately 29 km, originating near the towns of San Jose de Alianza, Nuevo Jerusalen, and La Rinconada Lajeña. The river then flows into the Marañon River at the Pedregales hamlet in the Bellavista district, within Jaen province, Cajamarca department [45].

A Digital Elevation Model (DEM) from the National Aeronautics and Space Administration (NASA) [46] was employed to determine that the maximum elevation within the subbasin reaches 3222 meters above sea level, while the lowest point is situated at 408 meters above sea level. The region is characterized by steep slopes, ranging from 40º to 56º in the upper and middle sections, and from 0º to 25º in the lower section. These topographical features, as shown in Fig. 3, suggest that the area is suitable for agricultural activities and the establishment of urban centers.

Moreover, this subbasin is a critical source of water supply for human consumption, serving as the primary provider for the cities of Jaen and Bellavista, as well as their associated agricultural valleys. The locations of the sampling stations are detailed in Table II.

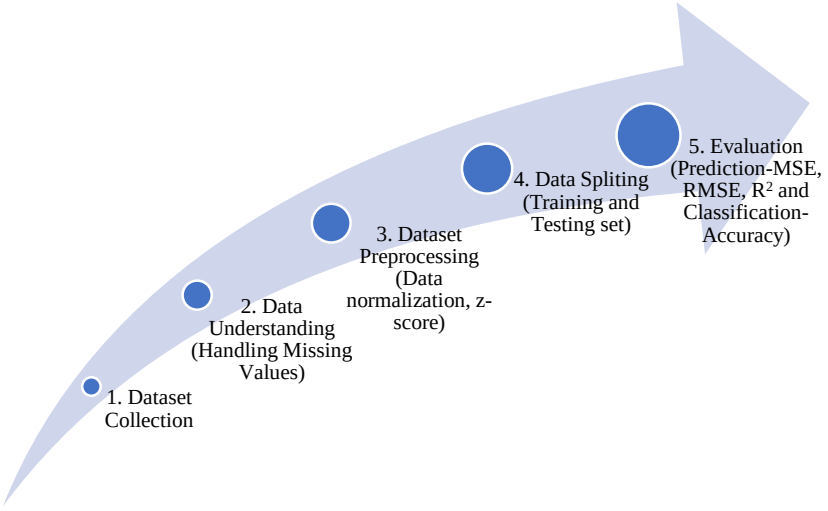

Fig. 1. Framework of the proposed methodology.

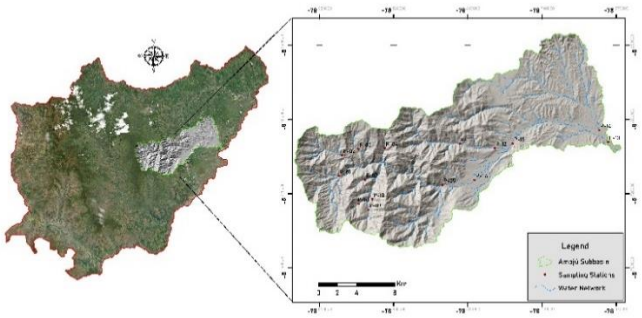

Fig. 2. Study area location map.

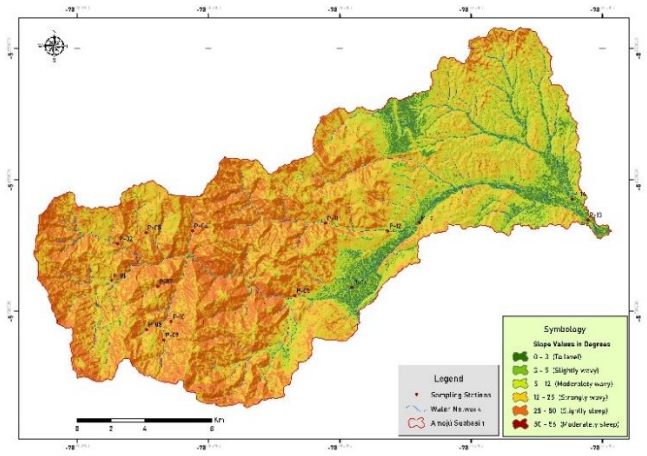

Fig. 3. A Subbasin slope map.

TABLE II. SAMPLING STATION LOCATIONS

Code	Location	Latitude	Longitude
P-01	Nuevo Jerusalén	-5.70347	-78.93161
P-02	La Rinconada Lajeña	-5.68396	-78.92851
P-03	San Antonio	-5.67808	-78.91332
P-04	La Cascarilla	-5.67719	-78.88849
P-05	Puente La Corona	-5.71176	-78.83388
P-06	Punte Pakamuros	-5.70729	-78.80355
P-07	Cruzpa Huasi	-5.70682	-78.90707
P-08	La Granadilla	-5.72972	-78.91294
P-09	La Virginia	-5.73523	-78.90385
P-10	La Victoria	-5.72548	-78.89995
P-11	Puente Tumbillán	-5.67307	-78.81757
P-12	Qda. Tumbillán- Altura La Granja	-5.6772	-78.78447
P-13	Puente Bellavista Viejo	-5.67135	-78.67758
P-14	Puente Santa Cruz	-5.66024	-78.68604
P-15	Yanuyacu - Altura UNJ	-5.67278	-78.7676

The upper section of the basin is dominated by a Tropical Premontane Rainforest (bh-PT) with annual precipitation levels reaching up to 1968 mm. The middle section is characterized by a very humid Tropical Low Montane Forest (bmh-MBT). Agroforestry systems, particularly those cultivating Coffea arabica (coffee) in association with citrus trees, are prevalent in the upper subbasin. The lower subbasin is dominated by a Tropical Premontane Dry Forest (bs-PT), where extensive rice fields (Orysa sativa) are cultivated.

B. Hydrogeological Settings

The water chemistry and quality in the study area influenced by both the lithology and the duration of water-rock interaction [47]. To identify the aquifer units within the National Geological Map provided by the Geological, Mining, and Metallurgical Institute (INGEMMET) [48] was used. This data was processed using QGIS software to create a hydrogeological map, illustrating the characteristics of the lithological units, as shown in Fig. 4.

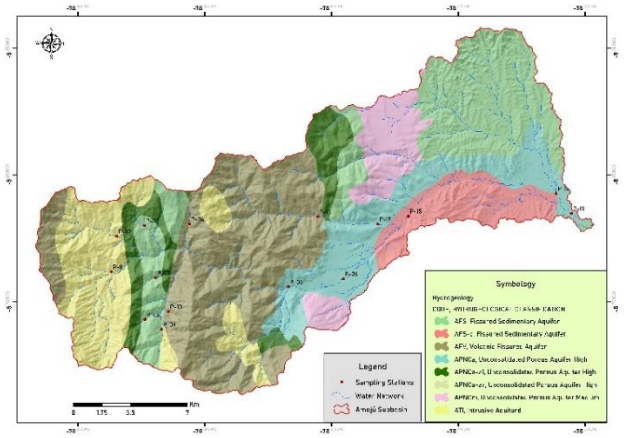

Fig. 4. Hydrogeolocgical map of the subbasin.

The lithological units within the subbasin display varied hydrogeological properties. Based on the hydrogeological map, these units were classified into aquifers and aquitards. Six distinct hydrogeological categories were identified within the aquifer units, while only one category was identified within the aquitards, as detailed in Table III.

C. Dataset Collection

Water samples were collected from 15 designated sampling points (Upper, Middle and Lower) within the Amoju Hydrographic Subbasin over a period from November 15th, 2023 to July 20th, 2024. The geographic distribution of these sampling points across the subbasin is illustrated in Fig. 5. Samples were brought to the CEASA and CAE laboratory in an insulated cooler box containing ice packs to maintain a stable temperature during transit. Analytical procedures commenced within 48 hours of sample collection to ensure data integrity.

The dataset for this study was sourced from strategic locations in Jaen-Cajamarca, focusing on five (05) physicochemical parameters measured in situ at each of fifteen (15) sampling sites. These parameters include pH, Electrical Conductivity (EC), Dissolved Oxygen (DO), total dissolved solids (TDS), and temperature (T°), all of which were measured using a recalibrated portable HANNA Multi-parameter HI 9829 device. Additionally, the dataset includes four (04) non-metallic inorganic parameters: alkalinity ($CaCO_3$), total hardness (TH), nitrates (NO_3^{1-}), and phosphates (PO_4^{3-}). Table IV provides a detailed description of each attribute measured.

TABLE III. A DESCRIPTION OF THE HYDROGEOLOGICAL CHARACTERISTICS OF THE SUBBASIN [48]

Hydrogeological Unit	Classification Hydrogeological	Code	Lithology	Description Hydrogeological
Aquifer	Volcanic Fissured Aquifer	AFV	Andesites and Dacitas	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Fissured Sedimentary Aquifer	AFS	Lutites, intercalated with limestones, marls	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Fissured Sedimentary Aquifer	AFS-c	Conglomerates, shales and sandstones	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Unconsolidated Porous Aquifer High	APNCa-al	Alternation of shales and sands	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer High	APNCa	Alluvial, moraines, glaciofluvial, lacustrine and travertine	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer High	APNCa-ar	Sands, sandstones, gravels and conglomerates	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer Medium	APNCm	Conglomerates, shales, mudstones	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquitard	Intrusive Aquitard	ATI	Acid and intermediate intrusive rocks	Formations without aquifers (with a very low permeability) can be considered.

TABLE IV. FEATURE DESCRIPTION OF DATASET

Attributes Name	Description
Physicochemical parameters	
pH	Water acidity and basicity.
EC	Water's ability to conduct electricity in the presence of ions.
DO	Concentration of dissolved oxygen in water.
TDS	Concentration of dissolved minerals, salts, metals, cations, or anions in water.
T°	Water temperature at the time of testing.
Non-metallic inorganic parameters	
$CaCO_3$	Water's ability to neutralize acids or resist changes that cause acidity, maintaining pH levels.
TH	Concentration of dissolved calcium and magnesium in water.
NO_3^{1-}	Concentration of nitrates, representing the most common form of nitrogen in water.
PO_4^{3-}	Concentration of phosphates, indicate of phosphorous and oxygen compounds in water.

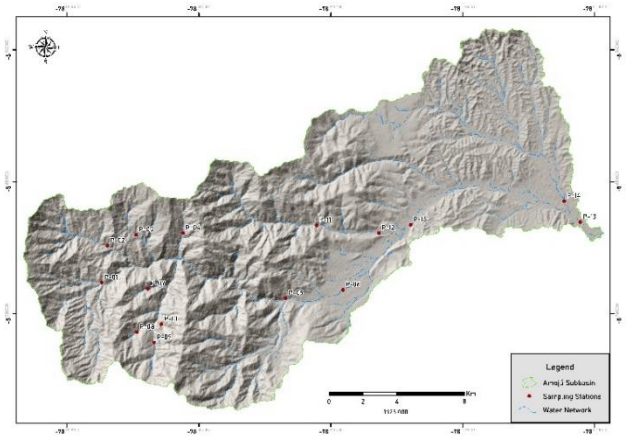

Fig. 5. Groundwater sampling sites.

D. Dataset Preprocessing and WQI Calculation

Data preprocessing is a critical step in ensuring the integrity and reliability of subsequent analysis. In this study, preprocessing involved the handling of missing data, normalization, and outlier removal to enhance the quality of the dataset. The Water Quality Index (WQI) was calculated to provide a comprehensive assessment of water quality across various sampling sites within the study area.

Water Quality Index (WQI): Methodology and Calculation

The Water Quality Index (WQI) is a widely recognized metric that synthesizes multiple water quality parameters into a single, overall score, reflecting the water's suitability for many uses [14]. The WQI in this study using the formula showing in Eq. (1).

$$WQI = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i}^{N} w_i} \quad (1)$$

Where N represents the number of parameters analyzed, q_i denotes the quality rating scale for each parameter i, computed in Eq. (2), and w_i denotes the unit weight of each parameter determined by Eq. (3).

$$q_i = 100 \times \left(\frac{V_i - V_{id}}{S_i - V_{id}}\right) \quad (2)$$

Where q_i is the parameter's actual value in the water samples tested, V_i represents the estimated value of the parameter i, V_{id} represents the ideal value under pure water conditions, and S_i represents the standard permissible limit for the parameter i as shown in Table V. The unit weight w_i is the parameter's recommended standard value as depicted in Table VI.

$$w_i = \frac{K}{S_i} \quad (3)$$

Where K denotes the proportionality constant, which is calculated using Eq. (4):

$$K = \frac{1}{\sum_{i=1}^{N} S_i} \quad (4)$$

The permissible limits and corresponding unit weights for the parameters are detailed in Table V and Table VI, respectively.

TABLE V. PERMISSIBLE LIMITS OF THE PARAMETERS USED IN CALCULATING THE WQI [49]

Parameter	Unit	Permissible Limits
pH	-	8.5
Conductivity	µS/cm	1000
Dissolved Oxygen	mg/L	10
Total Dissolved Solids	mg/L	1000
Temperature	°C	25
Alkalinity	mg/L	200
Hardness	mg/L	200
Nitrate	mg/L	45
Phosphate	mg/L	0.1

TABLE VI. PARAMETERS UNIT WEIGHTS

Parameter	Unit Weight (w_i)
pH	0.00004727
Conductivity	0.00000040
Dissolved Oxygen	0.00004018
Total Dissolved Solids	0.00000040
Temperature	0.00001607
Alkalinity	0.00000201
Hardness	0.00000201
Nitrate	0.00000893
Phosphate	0.00401830

The WQI is a versatile metric that can be employed for the calculation of numerous parameters, including those selected for analysis. The WQI depends on the variable data. The proposed system is capable of testing any parameters in conjunction with any water quality data.

IV. RESULTS AND DISCUSSION

Table VII provides descriptive statistics for the dataset attributes derived from 75 groundwater samples. These statistics, including count, mean, standard deviation, minimum, maximum, and quartiles, offer a comprehensive overview of the dataset's distribution and underlying properties. The mean pH value of 7.79 with a standard deviation of 0.46 suggests a slightly basic water quality, consistent with findings from similar studies in hydrographic subbasins [50], [51]. Conductivity averages at 220.96 µS/cm, reflecting significant variability in ion concentration among the samples. Dissolved Oxygen (DO) levels, averaging 7.46 mg/L, are critical for sustaining aquatic life, aligning with established benchmarks [52]. Total Dissolved Solids (TDS) display considerable variability with a mean of 164.61 mg/L, indicative of diverse mineral content in the samples. The mean temperature (Tº) of 21.52 ºC influences the solubility and reaction rates of various chemical constituents, further impacting water quality parameters [53].

Alkalinity and hardness, with means of 114.84 mg/L and 136.57 mg/L, respectively, reflect the water's buffering capacity and calcium/magnesium content, both of which are crucial for assessing the chemical stability of the water [25]. The mean concentrations of nitrate (NO_3^{1-}) and phosphate (PO_4^{3-}) are 0.021 mg/L and 1.52 mg/L, respectively, highlighting the presence of nutrient pollution, a significant concern in water quality management [50]. The Water Quality Index (WQI) has a mean value of 1.88, indicative of the overall quality of the water samples analyzed.

The correlation matrix presented in Fig. 6 is crucial for understanding the interrelationships among the water quality parameters. It enables the identification of functional dependencies, where strong correlations ($r > 0.7$) suggest significant associations, while weaker correlations ($r < 0.4$) imply more complex or indirect relationships. The WQI, the primary focus of this study, exhibits a strong positive correlation with phosphate levels ($r = 0.99$), underlining the significant impact of nutrient concentrations on overall water quality. In contrast, WQI shows weak correlations with parameters such as EC, TDS, $CaCO_3$, and NO_3^{1-}, suggesting that these variables, while influential, do not directly drive the WQI in this context.

The detailed examination of the correlations reveals that pH is moderately correlated with total hardness (TH), $CaCO_3$, and temperature (T°), with respective correlation coefficients of 0.59, 0.54, and 0.6. These findings are consistent with previous studies that have observed similar patterns in groundwater quality assessments [27]. Conductivity (EC) displays a strong positive correlation with T° ($r = 0.73$), $CaCO_3$ ($r = 0.94$), and TH ($r = 0.88$), indicating that these parameters are interdependent, likely due to their shared origin in mineral dissolution processes [14].

TABLE VII. DESCRIPTIVE STATISTICS OF THE FEATURES

Parameter	Count	Mean	Std Dev	Min	Q1	Median	Q3	Max
pH	75.00	7.790020	0.462099	6.75	7.4960	7.9340	8.1810	8.5510
Conductivity (µS/cm)	75.00	220.962667	197.740533	29.80	57.70	137.20	338.50	722.00
Dissolved Oxygen (mg/L)	75.00	7.4616	0.6773318	4.50	7.2250	7.60	7.8150	10.12
Total Dissolved Solids (mg/L)	75.00	164.605333	207.768171	6.00	38.65	95.90	216.50	1400.00
Temperature (°C)	75.00	21.52	5.052053	15.10	16.80	20.60	20.65	32.80
Alkalinity (mg/L)	75.00	114.84	86.908159	20.00	40.00	82.00	191.00	354.00
Hardness (mg/L)	75.00	136.5720	109.56334	34.20	39.90	102.60	225.15	427.50
Nitrate (mg/L)	75.00	0.021366	0.030259	0.000354	0.006854	0.010302	0.0240	0.218747
Phosphate (mg/L)	75.00	1.517073	1.220099	0.3215	0.862250	1.12130	1.6790	7.850
WQI	75.00	1.878745	1.183223	0.582148	1.238313	1.614984	2.089291	7.89930

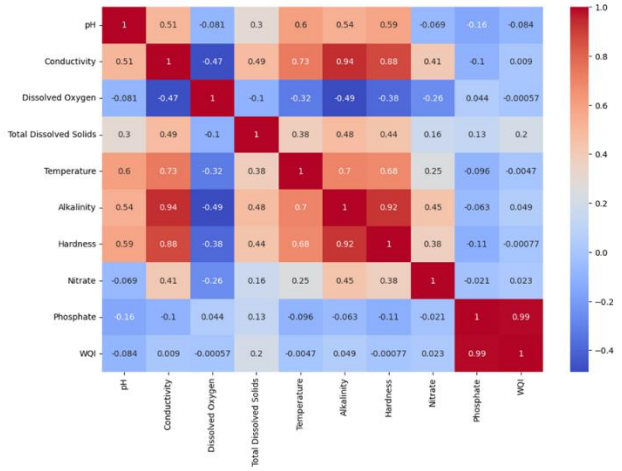

Fig. 6. Heatmap visualization of the future correlations.

Dissolved Oxygen (DO) exhibits a negative correlation with several parameters, notably EC (r = -0.47) and $CaCO_3$ (r = -0.49), which may suggest that higher ionic content and carbonate hardness could suppress oxygen solubility, a phenomenon that has been documented in other hydrographic contexts [50]. The analysis of TDS reveals moderate to strong positive correlations with pH (r = 0.3), EC (r = 0.49), and temperature (r = 0.38), reflecting the influence of these factors on dissolved solid concentrations [51]. The temperature itself strongly correlates with EC (r = 0.73) and $CaCO_3$ (r = 0.7), further reinforcing the interdependence of these water quality metrics.

The scatter plot matrix shown in Fig. 8 and the heatmap visualization provide additional insights into these relationships, offering a visual representation of the strength and direction of correlations. These graphical tools are essential for identifying patterns and anomalies in the dataset, facilitating a more nuanced interpretation of the results. The distribution of water compounds, as depicted in Fig. 7, confirms the trends observed in the correlation matrix, support, supporting the conclusion that physicochemical parameters, particularly nutrient concentrations, are critical determinants of water quality in the Amoju Hydrographic Subbasin.

The application of an Artificial Neural Network (ANN) model to predict the Water Quality Index (WQI) yielded robust results, demonstrating strong predictive performance across several key metrics, as shown in Table VIII. The model's Mean Absolute Error (MAE) of 0.2478 indicates a high degree of accuracy, with predicted WQI values deviating minimally from actual measurements. This level of accuracy is consistent with previous studies employing machine learning techniques for water quality prediction, further validating the model's effectiveness [44].

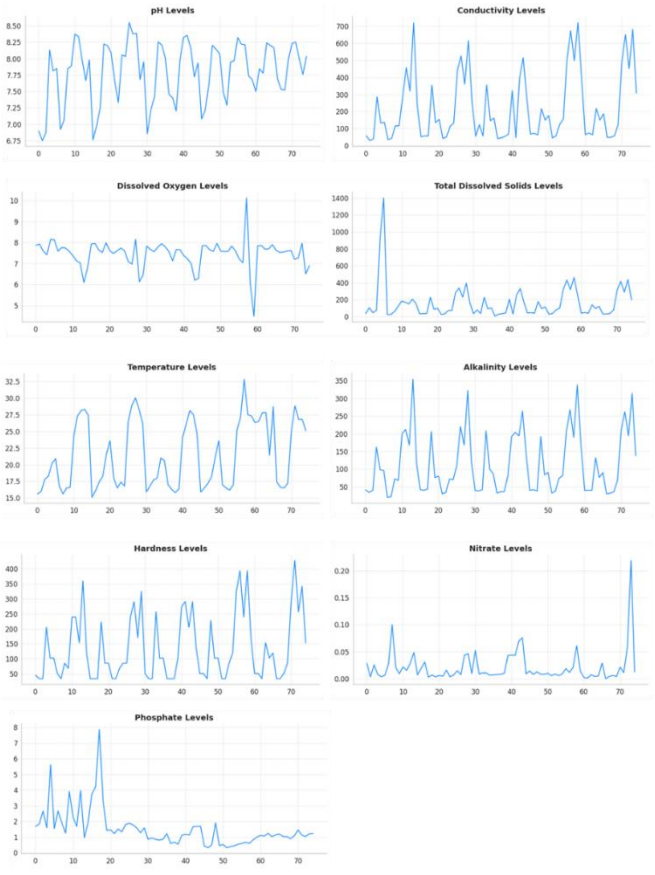

Fig. 7. Distribution of water compounds from the dataset.

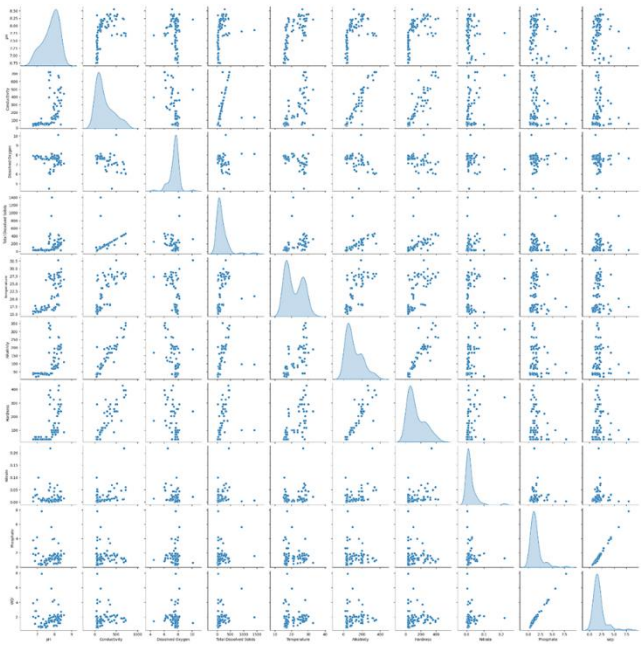

Fig. 8. Scatter plot matrix of the feature.

TABLE VIII. RESULTS BY ANN FOR WQI PREDICTION

Parameters	ANN
MAE	0.2478
MSE	0.0962
RMSE	0.3102
R2	0.9518

The model's Mean Squared Error (MSE) of 0.0962 and Root Mean Squared Error (RMSE), calculated as 0.3102 underscore the model's ability to minimize prediction errors, with the RMSE providing a direct measure of prediction in the same as the WQI. The model's predictive strength is further validated by the R-squared (R^2) value of 0.9518. The R^2 score suggests that approximately 95.18% of the variance in WQI can be explained by the model, highlighting its robustness and reliability as a predictive tool [23].

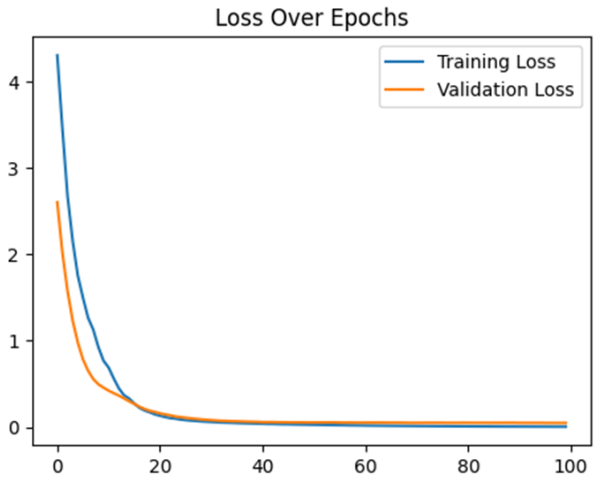

Fig. 9. ANN loss graph per epoch.

The convergence of the training and validation loss curves over 100 epochs, as illustrated in Fig. 9, suggest that the model is well-calibrated, with minimal risk of overfitting. This further corroborated by the "Actual vs. Predicted WQI" scatter plot shown in Fig. 10, which provides a visual comparison between the actual WQI values and those predicted by the model. The points on the plot are closely aligned along the red dashed line, which represents a perfect prediction. This close alignment reinforces the model's accuracy and the minimal prediction error. The plot confirms that the ANN model produces reliable predictions with a high degree of accuracy.

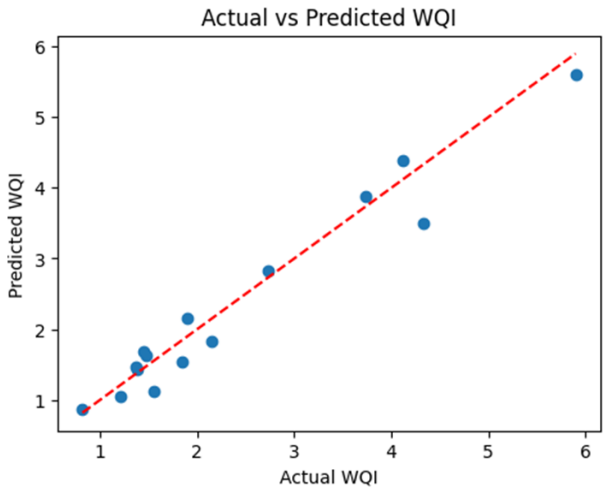

Fig. 10. Actual vs. predicted WQI graph.

V. CONCLUSION AND FUTURE WORK

This section presents a summary of the research conclusions and offers recommendations for future directions, highlighting key findings, limitations, implications, and suggests potential areas for further investigation.

A. Conclusion

Water is an essential resource for life on Earth, and ensuring its quality is fundamental to maintaining human health and environmental sustainability. The assessment of water quality, traditionally carried out through standard laboratory methods, has significantly advanced with the integration of machine learning (ML) techniques. These ML-based approaches offer a more robust and accurate means of predicting water quality indices by leveraging a wide array of water quality parameters.

This study has highlighted the relationships between various physicochemical parameters and the Water Quality Index (WQI) through a comprehensive analysis using Pearson's correlation matrix and Artificial Neural Network (ANN) model. The following key conclusions are drawn:

1) Correlation insights:

a) Very weak correlation: The WQI exhibits very weak correlations with parameters such as Electrical Conductivity (EC), Total Dissolved Solids (TDS), calcium carbonate ($CaCO_3$), and nitrate (NO_3^{1-}). These weak correlations suggest that these parameters, while part of the overall water quality

assessment, have limited direct influence on the WQI in this specific study context.

b) Strong correlation: A strong positive correlation is observed between WQI and phosphate (PO_4^{3-}), indicating that phosphate levels are critical determinant of water quality in the studied hydrographic subbasin.

c) Moderate correlations: The pH shows moderate correlation with Total Hardness (TH), $CaCO_3$, and Temperature (T°), implying a notable, though not dominant, role of these parameters in influencing the water quality. EC also displays a moderate correlation with pH, TDS, and NO_3^{1-}.

d) Negative correlation: EC is negatively correlated with Dissolved Oxygen (DO) and phosphate (PO_4^{3-}) , suggesting inverse relationships that may have implications for aquatic life and overall water chemistry.

2) Model performance: The ANN model developed for predicting the WQI demonstrated high predictive accuracy, validated by a low Mean Absolute Error (MAE), Mean Squared Error (MSE), and a high R-squared (R^2) value. These metrics confirm the model's ability to reliably predict water quality, making it a valuable tool for environmental monitoring and management.

3) Comparative analysis: The findings align with existing literature, reinforcing the importance of certain key parameters, particularly phosphate, in water quality assessments. The study not only corroborates the conclusions of previous research but also expands upon them by providing a nuanced understanding of parameter interrelationships within this specific geographical and environmental context.

B. Future Work

While the current study has provided significant insights into water quality prediction using machine learning, several avenues for future research remain:

1) Incorporation of additional parameters: Future studies should focus on incorporating additional chemical and biological parameters, such as heavy metals and microbial indicators, which might provide a more comprehensive water quality assessment.

2) Longitudinal data analysis: Extending the temporal scope of data collection over longer periods could allow for the examination of seasonal and climate variations in water quality, thereby enhancing the robustness of the predictive model.

3) Enhanced model architectures: Further refinement of the ANN model, or the application of more advanced machine learning techniques such as ensemble methods, deep learning, or hybrid models, could potentially improve predictive performance and uncover more complex relationships between parameters.

4) Geoespatial analysis: Integrating geospatial analysis tools with machine learning could provide spatially explicit predictions of water quality, which would be invaluable for regional water management and policy-making.

5) Real-time monitoring and prediction: Developing real-time monitoring systems, coupled with AI-driven predictive models, could facilitate the timely detection of water quality anomalies, enabling swift remedial actions and thereby safeguarding public health.

By addressing these areas in future research, the predictive capabilities and practical applications of machine learning in water quality management can be significantly advanced, contributing to more effective and sustainable environmental stewardship.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest. The manuscript has been reviewed and approved by all authors, who have no financial or personal relationships that could inappropriately bias or influence the content.

ACKNOWLEDGMENT

We gratefully acknowledge the Soil and Water Analysis Center of the National University of Jaen for their invaluable assistance with the water sample analyses. Our sincere appreciation extends to all the professionals whose expertise and insights significantly enriched the scope and rigor of this study.

REFERENCES

[1] Kar, Devashish, Wetlands and Lakes of the World. New Delhi: Springer India, 201d. C. Accedido: 5 de julio de 2024. [En línea]. Disponible en: https://library.wur.nl/WebQuery/titel/2074454.

[2] D. Kar, «Wetlands and their Fish Diversity in Assam (India)», Transylvanian Review of Systematical and Ecological Research, vol. 21, n.o 3, pp. 47-94, dic. 2019, doi: 10.2478/trser-2019-0019.

[3] N. Adimalla, «Groundwater Quality for Drinking and Irrigation Purposes and Potential Health Risks Assessment: A Case Study from Semi-Arid Region of South India», Expo Health, vol. 11, n.o 2, pp. 109-123, jun. 2019, doi: 10.1007/s12403-018-0288-8.

[4] S. Gaikwad, S. Gaikwad, D. Meshram, V. Wagh, A. Kandekar, y A. Kadam, «Geochemical mobility of ions in groundwater from the tropical western coast of Maharashtra, India: implication to groundwater quality», Environ Dev Sustain, vol. 22, n.o 3, pp. 2591-2624, mar. 2020, doi: 10.1007/s10668-019-00312-9.

[5] B. Dzwairo, Z. Hoko, D. Love, y E. Guzha, «Assessment of the impacts of pit latrines on groundwater quality in rural areas: A case study from Marondera district, Zimbabwe», Physics and Chemistry of the Earth, Parts A/B/C, vol. 31, n.o 15-16, pp. 779-788, ene. 2006, doi: 10.1016/j.pce.2006.08.031.

[6] T. A. Sinshaw, C. Q. Surbeck, H. Yasarer, y Y. Najjar, «Artificial Neural Network for Prediction of Total Nitrogen and Phosphorus in US Lakes», J. Environ. Eng., vol. 145, n.o 6, p. 04019032, jun. 2019, doi: 10.1061/(ASCE)EE.1943-7870.0001528.

[7] R. Barzegar y A. Asghari Moghaddam, «Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction», Model. Earth Syst. Environ., vol. 2, n.o 1, p. 26, mar. 2016, doi: 10.1007/s40808-015-0072-8.

[8] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, y A. Elshafie, «Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia», Neural Comput & Applic, vol. 28, n.o S1, pp. 893-905, dic. 2017, doi: 10.1007/s00521-016-2404-7.

[9] L. Xu y S. Liu, «Study of short-term water quality prediction model based on wavelet neural network», Mathematical and Computer Modelling, vol. 58, n.o 3-4, pp. 807-813, ago. 2013, doi: 10.1016/j.mcm.2012.12.023.

[10] W. C. Leong, A. Bahadori, J. Zhang, y Z. Ahmad, «Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)», International Journal of River Basin Management, vol. 19, n.o 2, pp. 149-156, abr. 2021, doi: 10.1080/15715124.2019.1628030.

[11] Y. Wu y S. Liu, «Modeling of land use and reservoir effects on nonpoint source pollution in a highly agricultural basin», J. Environ. Monit., vol. 14, n.o 9, p. 2350, 2012, doi: 10.1039/c2em30278k.

[12] A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, y R. N. Sankhua, «Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India», Model. Earth Syst. Environ., vol. 5, n.o 3, pp. 951-962, sep. 2019, doi: 10.1007/s40808-019-00581-3.

[13] P. Liu, J. Wang, A. K. Sangaiah, Y. Xie, y X. Yin, «Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment», Sustainability, vol. 11, n.o 7, p. 2058, abr. 2019, doi: 10.3390/su11072058.

[14] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, y M. Maashi, «Water Quality Prediction Using Artificial Intelligence Algorithms», Applied Bionics and Biomechanics, vol. 2020, pp. 1-12, dic. 2020, doi: 10.1155/2020/6659314.

[15] M. Hmoud Al-Adhaileh y F. Waselallah Alsaade, «Modelling and Prediction of Water Quality by Using Artificial Intelligence», Sustainability, vol. 13, n.o 8, p. 4259, abr. 2021, doi: 10.3390/su13084259.

[16] H. M. Mustafa, A. Mustapha, G. Hayder, y A. Salisu, «Applications of IoT and Artificial Intelligence in Water Quality Monitoring and Prediction: A Review», 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 968-975, ene. 2021, doi: 10.1109/ICICT50816.2021.9358675.

[17] S. Palabıyık y T. Akkan, «Evaluation of water quality based on artificial intelligence: performance of multilayer perceptron neural networks and multiple linear regression versus water quality indexes», Environ Dev Sustain, jun. 2024, doi: 10.1007/s10668-024-05075-6.

[18] M. Rezaie-Balf et al., «Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach», Journal of Cleaner Production, vol. 271, p. 122576, oct. 2020, doi: 10.1016/j.jclepro.2020.122576.

[19] T. Xu, G. Coco, y M. Neale, «A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning», Water Research, vol. 177, p. 115788, jun. 2020, doi: 10.1016/j.watres.2020.115788.

[20] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, y S. Kumar, «Prediction of groundwater quality using efficient machine learning technique», Chemosphere, vol. 276, p. 130265, ago. 2021, doi: 10.1016/j.chemosphere.2021.130265.

[21] S. V. Moghadam, A. Sharafati, H. Feizi, S. M. S. Marjaie, S. B. H. S. Asadollah, y D. Motta, «An efficient strategy for predicting river dissolved oxygen concentration: application of deep recurrent neural network model», Environ Monit Assess, vol. 193, n.o 12, p. 798, dic. 2021, doi: 10.1007/s10661-021-09586-x.

[22] J. Wu y Z. Wang, «A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory», Water, vol. 14, n.o 4, p. 610, feb. 2022, doi: 10.3390/w14040610.

[23] F. Rustam et al., «An Artificial Neural Network Model for Water Quality and Water Consumption Prediction», Water, vol. 14, n.o 21, p. 3359, oct. 2022, doi: 10.3390/w14213359.

[24] S. Dilmi y M. Ladjal, «A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques», Chemometrics and Intelligent Laboratory Systems, vol. 214, p. 104329, jul. 2021, doi: 10.1016/j.chemolab.2021.104329.

[25] Md. M. Hassan et al., «Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms»:, HCIS, vol. 1, n.o 3-4, p. 86, 2021, doi: 10.2991/hcis.k.211203.001.

[26] M. I. Khoirul Haq, F. Dwi Ramadhan, F. Az-Zahra, L. Kurniawati, y A. Helen, «Classification of Water Potability Using Machine Learning Algorithms», en 2021 International Conference on Artificial Intelligence and Big Data Analytics, oct. 2021, pp. 1-5. doi: 10.1109/ICAIBDA53487.2021.9689727.

[27] S. Kouadri, A. Elbeltagi, A. R. Md. T. Islam, y S. Kateb, «Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)», Appl Water Sci, vol. 11, n.o 12, p. 190, dic. 2021, doi: 10.1007/s13201-021-01528-9.

[28] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, y J. García-Nieto, «Efficient Water Quality Prediction Using Supervised Machine Learning», Water, vol. 11, n.o 11, p. 2210, oct. 2019, doi: 10.3390/w11112210.

[29] Z. Ahmad, N. A. Rahim, A. Bahadori, y J. Zhang, «Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks», International Journal of River Basin Management, vol. 15, n.o 1, pp. 79-87, ene. 2017, doi: 10.1080/15715124.2016.1256297.

[30] M. Sakizadeh, «Artificial intelligence for the prediction of water quality index in groundwater systems», Model. Earth Syst. Environ., vol. 2, n.o 1, p. 8, mar. 2016, doi: 10.1007/s40808-015-0063-9.

[31] H. Zare Abyaneh, «Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters», J Environ Health Sci Engineer, vol. 12, n.o 1, p. 40, dic. 2014, doi: 10.1186/2052-336X-12-40.

[32] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, y M. F. Ramli, «Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors», Marine Pollution Bulletin, vol. 64, n.o 11, pp. 2409-2420, nov. 2012, doi: 10.1016/j.marpolbul.2012.08.005.

[33] L. Wang et al., «Improving the robustness of beach water quality modeling using an ensemble machine learning approach», Science of The Total Environment, vol. 765, p. 142760, abr. 2021, doi: 10.1016/j.scitotenv.2020.142760.

[34] C. Sillberg, P. Kullavanijaya, y O. Chavalparit, «Water Quality Classification by Integration of Attribute-Realization and Support Vector Machine for the Chao Phraya River», J. Ecol. Eng., vol. 22, n.o 9, pp. 70-86, oct. 2021, doi: 10.12911/22998993/141364.

[35] M. Yilma, Z. Kiflie, A. Windsperger, y N. Gessese, «Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia», Model. Earth Syst. Environ., vol. 4, n.o 1, pp. 175-187, abr. 2018, doi: 10.1007/s40808-018-0437-x.

[36] S. I. Abba et al., «Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index», Environ Sci Pollut Res, vol. 27, n.o 33, pp. 41524-41539, nov. 2020, doi: 10.1007/s11356-020-09689-x.

[37] A. Elbeltagi, C. B. Pande, S. Kouadri, y A. R. Md. T. Islam, «Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India», Environ Sci Pollut Res, vol. 29, n.o 12, pp. 17591-17605, mar. 2022, doi: 10.1007/s11356-021-17064-7.

[38] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, y N. Kazakis, «Improving prediction of water quality indices using novel hybrid machine-learning algorithms», Science of The Total Environment, vol. 721, p. 137612, jun. 2020, doi: 10.1016/j.scitotenv.2020.137612.

[39] A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, y F. Sayyahi, «Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River)», KSCE J Civ Eng, vol. 22, n.o 7, pp. 2206-2213, jul. 2018, doi: 10.1007/s12205-017-1703-6.

[40] Y. Zhang et al., «Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network», Water Research, vol. 164, p. 114888, nov. 2019, doi: 10.1016/j.watres.2019.114888.

[41] S. B. H. S. Asadollah, A. Sharafati, D. Motta, y Z. M. Yaseen, «River water quality index prediction and uncertainty analysis: A comparative study of machine learning models», Journal of Environmental Chemical Engineering, vol. 9, n.o 1, p. 104599, feb. 2021, doi: 10.1016/j.jece.2020.104599.

[42] S. I. Khan, N. Islam, J. Uddin, S. Islam, y M. K. Nasir, «Water quality prediction and classification based on principal component regression and gradient boosting classifier approach», Journal of King Saud University - Computer and Information Sciences, vol. 34, n.o 8, pp. 4773-4781, sep. 2022, doi: 10.1016/j.jksuci.2021.06.003.

[43] D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, y N. T. D. Thuy, «Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam», Water, vol. 14, n.o 10, p. 1552, may 2022, doi: 10.3390/w14101552.

[44] M. Y. Shams, A. M. Elshewey, E.-S. M. El-kenawy, A. Ibrahim, F. M. Talaat, y Z. Tarek, «Water quality prediction using machine learning models based on grid search method», Multimed Tools Appl, sep. 2023, doi: 10.1007/s11042-023-16737-4.

[45] Autoridad Nacional del Agua, «Observatorio Nacional de Recursos Hídricos». Accedido: 9 de agosto de 2024. [En línea]. Disponible en: https://snirh.ana.gob.pe/onrh/MapaTematicoUH.aspx.

[46] NASA JPL, «NASADEM Merged DEM Global 1 arc second V001». NASA EOSDIS Land Processes Distributed Active Archive Center, 2020. doi: 10.5067/MEASURES/NASADEM/NASADEM_HGT.001.

[47] S. Varol y A. Davraz, «Evaluation of the groundwater quality with WQI (Water Quality Index) and multivariate analysis: a case study of the Tefenni plain (Burdur/Turkey)», Environ Earth Sci, vol. 73, n.o 4, pp. 1725-1744, feb. 2015, doi: 10.1007/s12665-014-3531-z.

[48] Instituto Geológico, Minero y Metalúrgico, «Mapa geológico del Perú», Repositorio Institucional INGEMMET, mar. 2023, Accedido: 6 de julio de 2024. [En línea]. Disponible en: https://repositorio.ingemmet.gob.pe/handle/20.500.12544/3837.

[49] [A. A. Al-Othman, «Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses», Arabian Journal of Chemistry, vol. 12, n.o 8, pp. 2104-2110, dic. 2019, doi: 10.1016/j.arabjc.2015.01.001.

[50] V. B. B Patil, S. M. Pinto, T. Govindaraju, V. S. Hebbalu, V. Bhat, y L. N. Kannanur, «Multivariate statistics and water quality index (WQI) approach for geochemical assessment of groundwater quality—a case study of Kanavi Halla Sub-Basin, Belagavi, India», Environ Geochem Health, vol. 42, n.o 9, pp. 2667-2684, sep. 2020, doi: 10.1007/s10653-019-00500-6.

[51] A. R. Md. T. Islam, N. Ahmed, Md. Bodrud-Doza, y R. Chu, «Characterizing groundwater quality ranks for drinking purposes in Sylhet district, Bangladesh, using entropy method, spatial autocorrelation index, and geostatistics», Environ Sci Pollut Res, vol. 24, n.o 34, pp. 26350-26374, dic. 2017, doi: 10.1007/s11356-017-0254-1.

[52] A. R. Md. T. Islam, M. T. Siddiqua, A. Zahid, S. S. Tasnim, y M. M. Rahman, «Drinking appraisal of coastal groundwater in Bangladesh: An approach of multi-hazards towards water security and health safety», Chemosphere, vol. 255, p. 126933, sep. 2020, doi: 10.1016/j.chemosphere.2020.126933.

[53] K. P. Singh, N. Basant, y S. Gupta, «Support vector machines in water quality management», Analytica Chimica Acta, vol. 703, n.o 2, pp. 152-162, oct. 2011, doi: 10.1016/j.aca.2011.07.027.