

Article

Comparative Performance Analysis of Machine Learning Models for Predicting the Weighted Arithmetic Water Quality Index

Bedia Çalış 1 , İbrahim Bayhan 1, Hamza Yalçın 2, İbrahim Öztürk 3,4 and Mehmet İrfan Yeşilnacar 1,* 

¹ Department of Environmental Engineering, Faculty of Engineering, Harran University, Şanlıurfa 63000, Türkiye; bediacalis@gmail.com (B.Ç.); ibayhan71@gmail.com (İ.B.)

² Department of Animal Science, Faculty of Agriculture, Harran University, Şanlıurfa 63000, Türkiye; hyalcin@harran.edu.tr

³ Department of Electrical and Electronics Engineering, Faculty of Engineering, Osmaniye Korkut Ata University, Osmaniye 80010, Türkiye; ibrahimozturk@osmaniye.edu.tr

⁴ Department of Electronics and Communication Engineering, Faculty of Technology, Karadeniz Technical University, Trabzon 61000, Türkiye

* Correspondence: mirfan@harran.edu.tr

Abstract

Precise water quality forecasting is vital for sustainable resource management and public health, especially in semi-arid environments. This study investigates the predictive capabilities of ten Machine Learning (ML) algorithms using a dataset of 308 drinking water samples collected from various districts in Şanlıurfa Province, Türkiye. We evaluated ten predictive models, including Support Vector Regressor (SVR) and Extreme Gradient Boosting (XGBoost), both integrated with dimensionality reduction and hyperparameter optimization. Nineteen physicochemical and microbiological parameters—Temperature, chlorine (Cl^-), pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS), nitrite (NO_2^-), nitrate (NO_3^-), ammonium (NH_4^+), sulfate (SO_4^{2-}), Free Chlorine (Cl_2), calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), fluoride (F^-), trihalomethanes (THMs), *Escherichia coli*, *Enterococci*, Total Coliform—were used as input features. The dataset was split into training (75%) and testing (25%) subsets, and model performance was assessed through 10-fold cross-validation and hold-out testing procedures. To improve model generalization and mitigate the effects of class imbalance, we implemented the Adaptive Synthetic Sampling (ADASYN) technique. ML algorithms were evaluated using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The LSTM model optimized using Randomized Search outperformed the SVR and XGBoost models, demonstrating the highest accuracy and generalization capability, as evidenced by the superior R^2 value of 0.999 following ADASYN balancing and the lowest RMSE (1.206). These findings underscore the effectiveness of the LSTM framework in modeling the complex variance of the Weighted Arithmetic Water Quality Index (WAWQI). The findings of this study are expected to support future water quality monitoring strategies, inform policy development, and contribute to sustainable water resource management in arid and semi-arid regions.

Keywords: machine learning; modeling; evaluation; WAWQI; LSTM



Academic Editor: Christos S. Akratos

Received: 3 February 2026

Revised: 4 March 2026

Accepted: 13 March 2026

Published: 16 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Access to drinking-quality water is a critical component of maintaining optimal human health. Modernization and industrial expansion have led to the introduction of diverse

anthropogenic contaminants into drinking water sources. The limited effectiveness of conventional treatment technologies, the production of toxic disinfection by-products (DBPs), and the instability of water distribution networks further exacerbate the risks to water quality and public health [1]. Additionally, the supply of quality groundwater is not limited to domestic and agricultural irrigation; good-quality groundwater is also used as a source of drinking water worldwide [2]. The excessive utilization of pesticides and fertilizers has the potential to compromise the integrity of aquatic ecosystems and the efficient usage of water resources [3]. In light of these difficulties, assessing the quality of the water has great importance as the first step in creating effective plans for managing water resources and putting pollution control measures into action [4]. Consequently, scientific evaluation of water quality is imperative for integrated and forward-thinking water resource management. The evaluation of water quality has traditionally entailed a comparison of multiple water parameters to standard levels that are recommended for various gradations. This approach offers a comprehensive overview of the state of the water environment [5].

Monitoring water quality and detecting sanitary concerns are critical for safeguarding populations from waterborne infections and devising effective prevention strategies. Global water pollution is a substantial hazard to human health, with the effect and severity of health consequences determined by factors such as the chemical composition of pollutants, exposure length, and pollutant concentration levels [6–9]. Although safe potable water management requires monitoring, collecting and analyzing water quality data, evaluating and synthesizing the data can be challenging and time-consuming. Globally, more than 68 water quality assessment instruments have been developed to evaluate water quality data under diverse hydro-environmental conditions. Among these instruments, the Water Quality Index (WQI) models are a notable example [10]. The WQI was first introduced in the 1960s by Horton (1965) [11]. Weighted Arithmetic Water Quality Index (WAWQI) is a special method developed by Brown et al. (1972) that calculates the purity of water by using parameter weights to assess water quality [12]. It offers a clear and practical summary of the overall status of a water body by converting diverse physical, chemical, and biological data into a single index value [13,14]. WAWQI models use aggregation functions to condense large water quality datasets, which vary over time and space, into a single value that reflects the overall water quality. They are highly valued by water management and supply organizations due to their ease of use and ability to simplify complex water quality data into a single, comprehensible measure [10].

WQI approaches have become a prevalent global method for evaluating and communicating water quality status, as they transform specific water quality indicator data into unitless values while retaining the crucial information from raw measurements [15]. The ease of use of the index is to provide a clear and interpretable assessment of water conditions by categorizing water quality into levels such as “excellent”, “good”, “poor”, “very poor” or “unsuitable” [16]. Conventional indexing methodologies are inadequate for processing large-scale and structurally complex datasets, particularly when timely assessments are required. Although almost all index models have gained widespread use in water quality assessment because of their relatively simple and easily relatable outputs [10], ambiguity and eclipsing are two major limitations of the models in general [17]. This limitation has prompted the search for more robust and dynamic water quality monitoring approaches. In urban environments, especially those affected by seasonal variability, rapid urbanization, and technological change, the adoption of modern data-driven water management and prediction methods has become increasingly important [18,19].

Hence, it has become a trend to use ML approaches as a tool to assist parameter selection. With the capacity to handle complicated datasets and identify patterns that

conventional approaches can miss, ML techniques have emerged as extremely useful instruments for controlling and forecasting environmental concerns like rainfall or water quality [20]. Researchers have studied how to predict water quality using different sets of input variables chosen through correlation analysis. A comparative study was conducted in [21] to examine the performance of linear regression against seven ML techniques. The models considered in this analysis were as follows: Artificial Neural Networks (ANNs), Self-Organizing Maps (SOMs), K-Nearest Neighbor (KNN), Support Vector Machines (SVMs), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Stochastic Gradient Boosting (SGB). The incorporation of supplementary input variables enhanced the performance of all models, with the exception of SOM. Several models, including SVM, RF, XGB, and SGB, attained R^2 values exceeding 0.99 [22] and estimated WQI using SVM, KNN, Ensemble Tree (ET), and Discriminant Analysis (DA) methods. Their results indicated that SVM was the most effective classifier for both raw and normalized datasets. The prediction accuracy for training data reached 90.8% with raw data and 89.2% with normalized data. Meanwhile, testing data showed prediction accuracies of 86.67% and 93.33% for raw and normalized datasets, respectively. Han et al. (2025) [19] employed advanced ML techniques such as Long-short Term Memory (LSTM), RF, Decision Tree (DT), and SVM to predict WQI with the aim of enhancing urban environmental management. This approach sought to tackle the pressing issues of waste management and water quality in urban areas where rapid urbanization intensifies environmental degradation and poses significant public health risks. Among the algorithms tested on training, testing, and validation datasets, LSTM consistently outperformed others by delivering superior prediction accuracy and precision, achieving R^2 values above 0.9964 and RMSE values of 0.0611, 0.0810, and 0.0754, respectively. Similarly, Elmotawakkil et al. (2025) [20] assessed ML methods for WQI and categorization prediction, including ANN, DT, SVM, XGBoost, and LSTM. With an accuracy range of 99.07–99.99% and an R^2 value of 0.9999, respectively, XGBoost and LSTM perform noticeably better than the other models according to the data. An evaluation of five different algorithms for WQI estimation in a related study [23] identified the SVM as the most effective model. This conclusion was drawn from its leading performance metrics, registering a 95.4% accuracy on standardized data and 88.9% on the test samples. The other algorithms considered in this comparative work were DT, KNN, DA, and ET.

To assess the spatial distribution of WQI values and classes across Indian river systems, Singh and Sharma (2024) [24] conducted modeling experiments using five algorithms, namely DT, RF, Gradient Boosted Trees (GBTs), ANN, and SVM. The results showed that GBT, RF, and ANN were able to produce accurate predictions using only two input variables, while GBT achieved over 80% performance metrics with just three input parameters. Similarly, Mo et al. (2024) [14] aimed to identify essential water quality parameters and the most reliable models for improving WQI estimation accuracy while accounting for seasonal variability in aquatic environments. Their findings indicated that increasing the number of input variables generally enhanced model performance, with the exception of the Self-Organizing Map (SOM). Among the evaluated models, tree-based ensemble approaches such as XGBoost and RF demonstrated the highest predictive accuracy. In another study, Talukdar et al. (2024) [25] focused on developing advanced ML tools for WQI assessment and model interpretability, specifically targeting lake water quality evaluation. The H2O API grid search algorithm was employed to optimize models including Deep Neural Networks (DNNs), RF, and Gradient Boosting Machines (GBMs). The analysis revealed that 52.7% of the samples were classified as extremely poor, corresponding to WQI values ranging from 72.38 to 100. Among the tested models, RF outperformed GBM and DNN, achieving higher R^2 values (0.97 for training and 0.90 for testing) along with lower RMSE, indicating superior accuracy and generalization capability. In water quality

research, multivariate statistical techniques are also widely applied. Principal Component Analysis (PCA), for example, reduces the risk of overfitting in the model, especially in large and heterogeneous datasets in semi-arid regions by eliminating redundant water quality parameters that are related to each other. PCA is commonly used to minimize data size and preserve the maximum amount of information while identifying dominant patterns and pollution sources in complex water quality datasets [26,27]. In this context, PCA can help reduce specific variables to identify indices that can define variations in water quality data with minimal information loss [28]. Therefore, PCA has been widely used in numerous studies to support water quality assessment and interpretation [29–33].

Despite the growing body of literature on ML-based water quality prediction, several critical research gaps persist. First, the majority of existing studies have evaluated either a single model family or a limited combination of algorithms, without systematically comparing the joint effects of dimensionality reduction and hyperparameter optimisation within the same modelling framework [15,21,22]. Second, while oversampling techniques such as SMOTE have been applied to address class imbalance in water quality classification tasks [21], the more adaptive ADASYN method—which generates synthetic samples proportionally to the local classification difficulty of minority instances—has received comparatively limited attention in continuous WAWQI prediction pipelines [34]. Third, studies integrating PCA-based dimensionality reduction with both Grid Search and Randomised Search optimisation strategies across SVR, XGBoost, and LSTM architectures simultaneously remain scarce in the context of drinking water quality assessment. Fourth, semi-arid regions such as Southeastern Türkiye present hydrogeochemical conditions—including elevated mineralisation, pronounced seasonal variability, and co-occurring physicochemical and microbiological stressors—that differ substantially from the riverine or lake-based systems predominantly studied in the existing literature [3,28]. These regional characteristics demand tailored modelling approaches that account for non-linear parameter interactions and highly skewed concentration distributions. The present study addresses these gaps by developing and evaluating ten integrated ML configurations applied to the Weighted Arithmetic Water Quality Index (WAWQI) in a semi-arid drinking water context, providing a comprehensive benchmark for model selection under conditions of high hydrogeochemical heterogeneity and class imbalance.

In this study, the evaluated models include a kernel-based SVR, a tree-based ensemble XGBoost, and a deep learning LSTM architecture, each further enhanced through dimensionality reduction (with PCA) and hyperparameter optimization (with Randomized Search and Grid Search) strategies to predict water quality by Arithmetic-weighted WQI at both regional and local scales. Support Vector Regressor (SVR) and XGBoost were selected due to their proven effectiveness in WAWQI prediction and their ability to model complex, non-linear relationships among multiple water quality parameters [35]. To further enhance model performance, dimensionality reduction using PCA and hyperparameter optimization through Grid Search and Randomized Search were incorporated. We incorporated a Long Short-Term Memory (LSTM) network to specifically capture the temporal dependencies and seasonal fluctuations inherent in the monitoring data [36]. This integrated modeling framework enables a comprehensive assessment of both spatial and temporal variations in water quality.

2. Materials and Methods

2.1. Study Area and Methodological Description

This study was conducted in Şanlıurfa Province (Figure 1), located in the southeastern part of Türkiye within the Southeastern Anatolia Region. The maps of the study area were produced from the topographic base maps of the region via ArcMap 10.8 software.

Şanlıurfa is geographically situated between $36^{\circ}38'$ and $38^{\circ}04'$ North latitudes and $37^{\circ}04'$ and $39^{\circ}00'$ East longitudes and covers an area of approximately $18,584 \text{ km}^2$, making it one of the largest provinces in the region. Şanlıurfa is characterized by a semi-arid continental climate [37]. Summers are extremely hot, with temperatures frequently exceeding 40°C and virtually no rainfall, placing the region among the hottest non-desert climates globally. Winters are generally cold and relatively wet, with occasional frosts and minimal snowfall. Spring and autumn represent transitional seasons, marked by relatively high temperatures and low precipitation levels [38]. The annual average temperature in the province is approximately 18.5°C , while mean annual precipitation is relatively low, averaging around 460 mm. The Euphrates River, one of the most significant water bodies in the Middle East, flows along the northern and eastern boundaries of the province. Water resources derived from the river play a critical role in regional agriculture through extensive irrigation networks developed under the Southeastern Anatolia Project (GAP) [39]. Topographically, Şanlıurfa is dominated by plateaus and plains, with elevations ranging from 400 to 800 m above sea level. In recent decades, rapid urbanization and population growth have increased pressure on natural resources, particularly drinking water supply and quality. The pronounced climatic extremes, intensive agricultural irrigation practices, and increasing urban pressure on water resources make Şanlıurfa Province a strategically relevant case study for evaluating drinking water quality and its spatial-temporal variability using integrated water quality indices and data-driven modeling approaches.

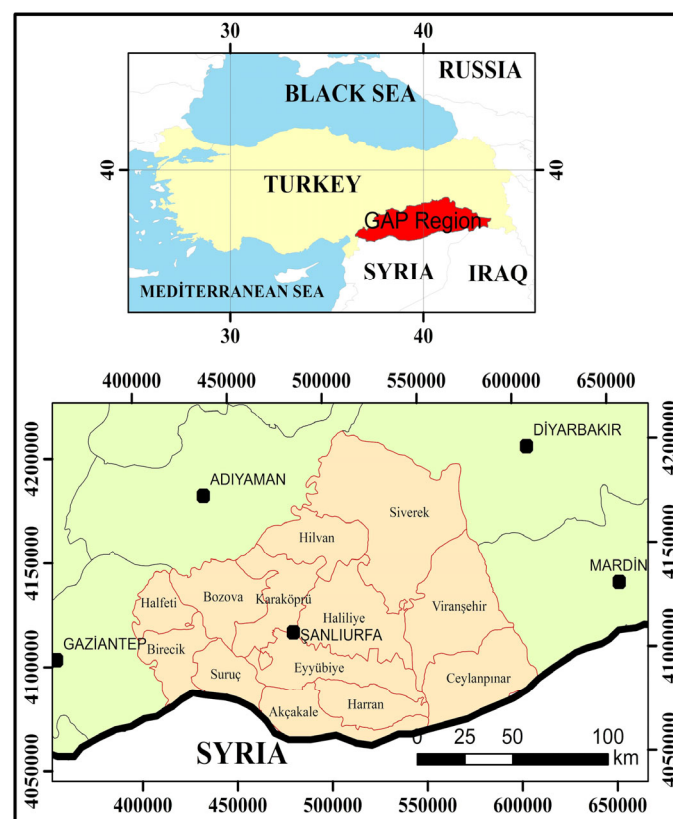


Figure 1. Study Area.

The study was structured into five main phases, as illustrated in Figure 2: (i) data collection and characterization, (ii) data preprocessing, (iii) dimensionality reduction and hyperparameter optimization, (iv) ML modeling, and (v) performance evaluation. To predict drinking water quality in 13 district centers of Şanlıurfa, a total of 308 drinking water samples were collected from 77 sampling points during January–September 2018.

Samples were taken from active fountains and taps directly connected to the water network at monitoring points designated by the Provincial and District Health Directorates. Groundwater serves as the main drinking water source in several districts. Siverek, Viranşehir, and Halfeti each have one spring, while Akçakale, Harran, Ceylanpınar, and Bozova rely on deep boreholes. In the Hilvan district, the primary water source is the Atatürk Dam Lake. Raw water is conveyed through a 17 km pipeline to a 1500 m³ reservoir. It is then transferred to a second, adjacent 1500 m³ reservoir for chlorination before being released into the public distribution network. Notably, the district lacks a formal drinking water treatment plant. The Suruç and Birecik districts utilize caisson wells drilled within the water reserve areas of the Birecik and Karkamış Dam Lakes. Additionally, Suruç benefits from five deep boreholes located in the district center. Birecik also draws from springs located within its streets. In Akçakale and Ceylanpınar, some wells are directly connected to the mains without an intermediate reservoir. The Siverek district also utilizes 28 springs in addition to the Karacadağ springs, with key sources. Similarly, in Viranşehir, 26 of the 40 springs are directly integrated into the public water supply system without a reservoir. Nineteen parameters—Temperature, Cl[−], pH, EC, TDS, NO₂[−], NO₃[−], NH₄⁺, SO₄^{2−}, Cl₂, Ca²⁺, Mg²⁺, Na⁺, K⁺, F[−], THMs, *E. coli*, *Enterococci*, Total Coliform—were selected in accordance with the national drinking water standards (ITASHY). Elemental analyses were conducted following the USEPA Method 200.7, while physicochemical parameters were measured according to ITASHY protocols and microbiological parameter measurements were performed via membrane filter method. The dataset was divided into 75% training and 25% testing subsets, a commonly adopted split in predictive modeling studies, to train the algorithms and evaluate their predictive performance on previously unseen data. ML models applied in this study included SVR, SVR with Principal Component Analysis (SVR + PCA), SVR with Grid Search (SVR + GS), SVR with Random Search (SVR + RS), XGBoost, XGBoost + PCA, XGBoost + GS, XGBoost + RS, LSTM, and LSTM + RS. Models were grouped according to their purpose: baseline models (SVR, XGBoost, LSTM), dimensionality-reduced models (SVR + PCA, XGBoost + PCA), and hyperparameter-optimized models (SVR + GS/RS, XGBoost + GS/RS, LSTM + RS). The seasonal structure of our dataset enabled the LSTM to identify time-varying patterns that stationary models often overlook.

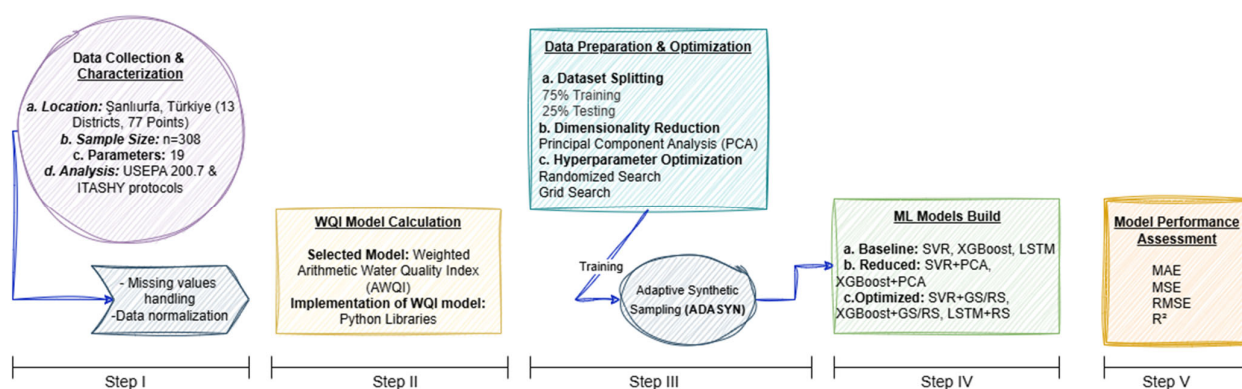


Figure 2. The methodological framework of the study.

2.2. Dataset Description

A total of 308 water samples were collected from 77 sampling points (Table 1) across 13 districts of Şanlıurfa province between January–September 2018. The analyzed parameters included Temperature, Cl[−], pH, EC, TDS, NO₂[−], NO₃[−], NH₄⁺, SO₄^{2−}, Cl₂, Ca²⁺, Mg²⁺, Na⁺, K⁺, F[−], THMs, *E. coli*, *Enterococci*, and Total Coliform [40]. Following the protocols established by the District Health Directorate, our sampling strategy involved collecting

drinking water from active public dispensers. The samples were taken in accordance with the methods determined in the “Handbook on Sample Collection, Transport and Analysis of Water for Human Consumption” published by the Ministry of Health in 2008. Microbiological samples were taken in 500 mL pure thiosulfate polypropylene bottles, and physicochemical samples were taken in 1.5 L polyethylene bottles. All selected points were confirmed to be directly connected to the water supply network and were part of the directorate’s official monitoring system. All procedures for the collection and conveyance of water samples were executed in accordance with the “2018 Criteria for Acceptable Water Specimens.” This regulatory standard was set forth by the Office of Public Health and Consumer Safety, functioning within the General Directorate of Public Health of the Ministry of Health [41].

Microbiological analyses were carried out at ŞUSKİ (General Directorate of Water and Sewerage Administration of Şanlıurfa Metropolitan Municipality) Drinking Water Treatment Plant Laboratory with the financial support of HUBAP (The Coordinating Unit for Scientific Research Projects at Harran University, Project No: 17198). Total Coliform, *E. coli* (TS EN ISO 9308-1), and *Enterococci* (TS EN ISO 7899-2) measurements were performed via membrane filter method [42,43]. The water samples were gently shaken to ensure homogeneity. A 100 mL water sample was filtered from three membrane filters prepared for Total Coliform, *E. coli* and *Enterococci*. After filtering process, Total Coliform and *E. coli* were incubated in chronogeneous agar broth at 37 °C for 24 h, whereas *Enterococci* was incubated in Avida broth at 37 °C for 48 h. Colonies formed at the end of the incubation period were counted. It was observed that red colonies were formed for Total Coliform, blue for *E. coli* and cherry-colored colonies for *Enterococci*. Temperature, pH, EC, and Cl_2 were measured in situ during sample collection. Temperature was recorded using a digital thermometer by immersing the probe into the water until the display reading stabilized. pH and EC measurements were performed using a Hach HQ14D Digital Conductivity Meter, while Cl_2 levels were determined colorimetrically on-site using a Hach device. TDS values were calculated by multiplying the EC values by a conversion coefficient ($k = 0.65$). Laboratory analyses for NO_2^- , NO_3^- , SO_4^{2-} , NH_4^+ , Cl^- , and F^- were conducted spectrophotometrically. The spectrometer was calibrated to the specific wavelength of each parameter before measuring concentration values. For THMs, Ca^{2+} , Mg^{2+} , Na^+ , and K^+ analysis was performed via Inductively Coupled Plasma Optical Emission Spectrometry (ICP-OES) following EPA Method 200.7. All analytical results were evaluated in accordance with the Turkish Regulation on Water Intended for Human Consumption (ITASHY) and EPA standards [44].

Table 1. Sampling Points and Coordinates.

Latitude	Longitude	District	Points No	Latitude	Longitude	District	Points No
4178974	529640	Siverek	42	4063454	494474	Akçakale	1
4179010	531017	Siverek	43	4062981	493992	Akçakale	2
4178327	530306	Siverek	44	4062973	495412	Akçakale	3
4177804	529095	Siverek	45	4062473	496230	Akçakale	4
4176927	527720	Siverek	46	4063306	496055	Akçakale	5
4093117	448909	Suruç	47	4097251	408987	Birecik	6
4092556	448595	Suruç	48	4096653	409847	Birecik	7
4090569	440580	Suruç	49	4099137	409637	Birecik	8
4091832	448471	Suruç	50	4098954	409340	Birecik	9

Table 1. Cont.

Latitude	Longitude	District	Points No	Latitude	Longitude	District	Points No
4091677	449623	Suruç	51	4098939	407399	Birecik	10
4092743	449160	Suruç	52	4098628	407791	Birecik	11
4121534	566565	Viranşehir	53	4135457	458524	Bozova	12
4122352	568473	Viranşehir	54	4135378	457742	Bozova	13
4121752	567968	Viranşehir	55	4134898	457455	Bozova	14
4121111	567754	Viranşehir	56	4135259	456916	Bozova	15
4120585	568308	Viranşehir	57	4135141	458092	Bozova	16
4120100	566825	Viranşehir	58	4135623	458212	Bozova	17
4119580	568474	Viranşehir	59	4078082	593457	Ceylanpınar	18
4120569	567184	Viranşehir	60	4077841	593602	Ceylanpınar	19
4122735	567735	Viranşehir	61	4079309	593795	Ceylanpınar	20
4106582	487150	Eyyübiye	62	4079838	593546	Ceylanpınar	21
4109418	481375	Eyyübiye	63	4080810	592712	Ceylanpınar	22
4111736	481819	Eyyübiye	64	4081934	592020	Ceylanpınar	23
4112027	480260	Eyyübiye	65	4123036	399517	Halfeti	24
4109980	477405	Eyyübiye	66	4122498	399716	Halfeti	25
4114340	484400	Eyyübiye	67	4120809	406323	Halfeti	26
4113634	682107	Haliliye	68	4120460	406197	Halfeti	27
4113623	480300	Haliliye	69	4121307	406964	Halfeti	28
4112314	483520	Haliliye	70	4079129	503012	Harran	29
4108791	486594	Haliliye	71	4079413	502882	Harran	30
4114173	483880	Haliliye	72	4079378	503195	Harran	31
4115237	482543	Karköprü	73	4080690	502214	Harran	32
4116705	482706	Karköprü	74	4159883	496032	Hilvan	33
4121051	482562	Karköprü	75	4160247	496319	Hilvan	34
4123427	480924	Karköprü	76	4161565	494296	Hilvan	35
4119209	480061	Karköprü	77	4155678	496192	Hilvan	36
				4159177	496546	Hilvan	37
				4177018	527299	Siverek	38
				4178521	526640	Siverek	39
				4179685	527716	Siverek	40
				4179306	528395	Siverek	41

The reliability of the analytical dataset was ensured through several quality assurance measures. In situ instruments—including the Hach HQ14D conductivity meter and the colorimetric chlorine analyser—were factory-calibrated and recalibrated prior to each sampling campaign in accordance with the manufacturer’s protocols. Spectrophotometric analyses were performed following wavelength-specific instrument calibration using certified standard solutions for each target analyte. ICP-OES measurements for THMs, Ca²⁺, Mg²⁺, Na⁺, and K⁺ were carried out following EPA Method 200.7, which prescribes multi-point external calibration and internal standard correction to minimise matrix interferences.

Microbiological analyses adhered to TS EN ISO 9308-1 for *E. coli* and Total Coliforms and TS EN ISO 7899-2 for *Enterococci*, both of which are internationally harmonised standards incorporating positive and negative control requirements. All laboratory analyses were conducted at the Şanlıurfa Water and Sewerage Administration Drinking Water Laboratory, which operates under the regulatory oversight of the General Directorate of Public Health [41,44]. It is acknowledged that the absence of reported duplicate sample recovery rates and certified reference material (CRM) data constitutes a limitation of the present study; these elements should be incorporated into future monitoring campaigns to provide a more complete characterisation of analytical uncertainty in data-driven water quality modelling frameworks [16,17].

2.3. Machine Learning Algorithms

ML methods have been extensively applied to fundamental problem types, including classification, regression, and clustering. In particular, classification techniques are widely preferred in water quality assessment due to their ability to provide robust performance indicators. The utilization of statistical analysis in examining historical water quality parameters facilitates the development of classification models that demonstrate high accuracy in categorizing the health status of water bodies. These approaches have been shown to contribute significantly to the prediction and monitoring of water quality indices, thereby enabling the development of resource management and early warning systems. In this study, a suite of ML algorithms, specifically SVR, SVR + PCA, SVR + GS, SVR + RS, XGBoost, XGBoost + PCA, XGBoost + GS, XGBoost + RS, LSTM, and LSTM + RS, were employed to predict water quality. These configurations were selected to evaluate the impact of dimensionality reduction and hyperparameter optimization on predictive accuracy. The machine learning and deep learning architectures described in this study were implemented using the Python (version 3.12.3) programming language, specifically leveraging libraries such as Scikit-learn for SVR and hyperparameter optimization, XGBoost for gradient boosting, and TensorFlow/Keras for the LSTM framework. All data preprocessing, including the ADASYN oversampling and Winsorization, was performed via custom scripts based on standard numerical libraries to ensure the reproducibility of the results.

2.3.1. Support Vector Regressor (SVR)

SVR is a kernel-based method derived from Vapnik's statistical learning framework [45]. SVR estimates a continuous target by fitting a function within an ε -insensitive tube, penalizing only deviations outside this tolerance, which yields a "flattest" possible fit that generalizes well. This approach uses kernel functions to handle non-linear patterns, leading to a sparse solution controlled by support vectors and margin width. Notably, SVR's computational complexity does not depend on input dimensionality, and it often achieves high predictive accuracy. SVR has been widely applied for real-valued function estimation in environmental data [46]. The model defines an ε -insensitive tube around the regression function, where deviations smaller than ε are ignored. The final architecture was optimized using Grid Search to determine the regularization parameter C and the kernel bandwidth γ .

- SVR with PCA: A hybrid SVR model using PCA is constructed to reduce feature dimensionality before regression [47]. The training data are first scaled and then projected onto principal components explaining 95% of variance, yielding 14 components out of the original feature set [48,49]. These components (rather than the original variables) are used as inputs to the SVR training to build the regression model. By compressing the input space, PCA helps avoid overfitting and reduces noise, often

improving efficiency and robustness. In practice, this combined PCA–SVR approach preserves most information while accelerating training and can enhance prediction accuracy during test phase.

- **SVR with Grid Search:** To optimize SVR hyperparameters, a grid-search strategy with cross-validation is employed. A five-fold CV scheme is used to evaluate each candidate, and model performance is quantified by the R^2 . The search grid covered combinations of the regularization parameter C , kernel bandwidth γ , and kernel type (linear or RBF), totaling 18 parameter sets. For each set, SVR is trained on five folds (yielding 90 models in total), and the mean R^2 is recorded. The parameter combination with the highest average R^2 is selected for the final model. In scikit-learn, this process can be automated by GridSearchCV, which systematically evaluates parameter grids via CV. This objective search ensures the final SVR model uses the best settings found in the validation process [50].
- **SVR with Randomized Search:** An alternative hyperparameter search using randomized sampling is also implemented. Random search explores a wide hyperparameter space more efficiently than exhaustive grid search [51]. In this setup, 30 random combinations of SVR parameters were drawn from predefined distributions (e.g., $C \in [0.1, 1, 10, 100]$, $\epsilon \in [0.01, 0.1, 0.2]$, kernels $\in \{\text{linear, RBF, polynomial}\}$, and $\gamma \in \{\text{scale, auto}\}$). As with the grid search, each of these 30 combinations is evaluated by a five-fold CV using negative MAE as the score. This yielded 150 trained models. The combination yielding the lowest average MSE is selected for the final SVR configuration. In practice, randomized search often finds near-optimal hyperparameters with far fewer iterations than grid search, making it well-suited for large or complex parameter spaces. The tuned SVR is then applied to the test set for prediction.

2.3.2. Extreme Gradient Boosting (XGBoost)

XGBoost represents an optimized application of gradient boosting methods based on DT [52]. The incorporation of L_1 and L_2 regularization terms serves to regulate model complexity, thereby mitigating the risk of overfitting. The advanced algorithms employed to determine splitting points have two primary benefits. First, they accelerate the training process. Second, they improve computational efficiency. The XGBoost algorithm has been demonstrated to exhibit high performance in regression and classification tasks, a feat attributable to its capacity to process disparate data types and its parallel processing optimizations. Furthermore, it is widely utilized across a diverse array of applications due to its adept management of missing values and its capacity to customize loss functions.

- **XGBoost with PCA:** A hybrid model combining PCA and XGBoost is created to leverage both dimension reduction and powerful ensemble learning. As before, the data are scaled and reduced via PCA to 14 principal components explaining 95% of the variance. These components served as input features for an XGBoost regressor. XGBoost is an efficient implementation of gradient-boosted DT that incorporates both L_1 and L_2 regularization to control model complexity [53]. The model is trained on the PCA-transformed training set, and its generalization ability is tested by making predictions on the similarly transformed test set. The performance of the XGBoost + PCA model is quantified using standard regression metrics (R^2 , MAE, RMSE) and compared against other methods. By reducing dimensionality first, this pipeline can reduce noise and computational cost while maintaining predictive accuracy [48,49].
- **XGBoost with Grid Search:** XGBoost is an advanced gradient boosting algorithm that builds an ensemble of regression trees in a stage-wise manner. It extends traditional boosting by including optimized search for best split points and by incorporating regularization (L_1/L_2 penalties) to prevent overfitting. This regularized framework

accelerates training and improves generalization. XGBoost also supports sparse data and employs system-level optimizations (like cache-aware access and data compression) to scale to large datasets. Its flexible design allows it to handle various data types, and it has demonstrated state-of-the-art performance on many regression and classification tasks [53]. In this study, the final XGBoost model was constructed using the best hyperparameters and then run on the test set to record its prediction results.

- XGBoost with Randomized Search: To further optimize XGBoost's hyperparameters, a randomized search approach was applied. A wide parameter space (including `n_estimators` [100–500], `learning_rate` [0.03–0.3], `max_depth` [3–7], `subsample` [0.8–1.0], `colsample_bytree` [0.8–1.0], and `gamma` [0–0.2]) is defined, and 30 random combinations were sampled. Each sampled configuration was evaluated via five-fold CV using negative MSE as the metric. In total, 150 XGBoost models were trained and validated. The combination yielding the lowest CV-MSE is chosen to configure the final XGBoost model. This model is then trained on the full training data and used to predict on the test set. Randomized search accelerates hyperparameter tuning and often finds good solutions in fewer iterations than grid search [51].

2.3.3. Long-Short Term Memory (LSTM)

LSTM is a deep learning model based on Long Short-Term Memory that is also developed for time-series prediction. Input features and the target variable were first scaled to [0, 1] using Min–Max scaling. The data were then reshaped into a 3D tensor of shape (samples, time steps, features) to fit the LSTM layer requirements. The network architecture consisted of a single LSTM layer with 64 units (using the tanh activation function), followed by a dense output layer for regression. The model was compiled with the Adam optimizer and MAE loss. Training was run for 50 epochs with batch size 16, and performance was monitored on a validation split after each epoch. Early stopping or similar checks were used to guard against overfitting. LSTM networks were specifically developed to model temporal dependencies within sequential datasets [54], making them particularly effective for forecasting tasks involving time-series data.

The selection of the LSTM architecture is justified by its inherent capacity to model complex non-linearities and temporal dependencies within environmental datasets. While traditional regressors often fail to account for the sequential nature of seasonal water quality fluctuations, LSTM utilizes memory cells to retain information across time steps. This is particularly critical for the Şanlıurfa dataset, which exhibits high variance in parameters like Fluoride and THMs that are influenced by cumulative climatic stressors.

In the proposed LSTM framework with randomized search, random search was selected for hyperparameter optimization due to its ease of implementation, natural suitability for parallel computation, and its ability to sample the hyperparameter space more uniformly, thereby enabling a more reliable evaluation of hyperparameter importance [55]. Recent studies confirm that randomised hyperparameter search consistently identifies near-optimal configurations for deep learning architectures applied to water quality prediction tasks [20,21]. To address potential constraints related to the dataset size of 308 observations, we employed the ADASYN technique to adaptively balance classes and utilized a single-layer LSTM configuration. This approach minimizes the number of trainable parameters relative to the available data, a strategy confirmed by the stable convergence of training and validation loss curves.

2.4. Model Development

2.4.1. Input Data

The present study employs a comprehensive dataset comprising physicochemical parameters and water quality indices collected from diverse sampling locations. Raw measurements—including Temperature (°C), Cl₂ (mg/L), pH, EC (μS/cm), TDS (mg/L), NO₂[−] (mg/L), NO₃[−] (mg/L), NH₄⁺ (mg/L), SO₄^{2−} (mg/L), Cl[−] (mg/L), Ca²⁺ (mg/L), Mg²⁺ (mg/L), Na⁺ (mg/L), K⁺ (mg/L), F[−] (mg/L), THMs (μg/L), *E. coli*, *Enterococci* and Total Coliform—were initially stored in an Excel workbook. Each record was annotated with unique identifiers (ID, sample number) and categorical descriptors (season) to facilitate downstream analyses.

2.4.2. Data Processing

Prior to model development, data preprocessing included column renaming, type conversion, outlier treatment via Winsorization, encoding, and feature scaling. Column names were standardized by stripping whitespace and mapping to short labels (e.g., “Temp” to temperature_c) for consistency. Extreme values beyond the 1st and 98th percentiles were capped to mitigate the influence of outliers using a custom Winsorization function. Categorical district labels were transformed into numeric codes through LabelEncoder, while continuous predictors were standardized with StandardScaler to ensure zero mean and unit variance for all features prior to training. In addition to these preprocessing steps, class imbalance was addressed using the Adaptive Synthetic Sampling (ADASYN) method. ADASYN improves upon traditional oversampling by adaptively generating more synthetic samples for minority instances that are harder to classify, based on their local distribution. This approach emphasizes difficult regions in the feature space, thereby refining decision boundaries and reducing bias toward majority classes. The algorithm determines the number of synthetic samples for each minority instance using its difficulty ratio (r_i):

$$r_i = \frac{\text{majority neighbors of } x_i}{k} \quad (1)$$

where x_i represents the i^{th} minority class sample, and k is the number of nearest neighbors. The normalized weights G_i are computed as:

$$G_i = G \cdot \frac{r_i}{\sum_i r_i}, G = N_M - N_m \quad (2)$$

where G is the total number of synthetic samples to generate, N_M and N_m are the sizes of majority and minority classes. For each minority instance x_i , G_i synthetic samples are created by interpolation:

$$x_{\text{new}} = x_i + \delta \times (x_{zi} - x_i), \delta \sim U(0, 1) \quad (3)$$

This adaptive mechanism ensures that oversampling focuses on challenging regions of the feature space, thereby improving model generalisation in highly imbalanced datasets [34]. It is important to note that whilst ADASYN is conventionally associated with classification tasks, its application within the present regression-oriented pipeline is methodologically justified by the structure of the WAWQI target variable. Although WAWQI is treated as a continuous output variable throughout model training and evaluation, the underlying distribution of samples across the five quality classes—Excellent, Good, Poor, Very Poor, and Unsuitable—is markedly skewed, with the majority class (Good, 53.2%) outnumbering the minority classes (Very Poor and Unsuitable, 2.3% each) by a factor exceeding 23 (Table 2). Under such conditions, regression models trained on the

raw dataset are prone to systematic bias, disproportionately optimising predictions for the dominant quality range whilst underperforming across critical low-frequency classes that are of greatest public health concern [17,56].

Table 2. Distribution of WAWQI classes for all samples.

Category	Before ADASYN		After ADASYN	
	Count	Percent	Count	Percent
Excellent	118	38.3	164	20
Good	164	53.2	164	20
Poor	12	3.9	164	20
Very Poor	7	2.3	164	20
Unsuitable	7	2.3	164	20
Total	308	100%	820	100%

To address this limitation, ADASYN was applied to the discretized WAWQI class labels in a pre-processing step prior to regression model training. Specifically, each continuous WAWQI observation was first assigned to its corresponding quality class according to the classification thresholds defined by Brown et al. (1972) [12]. Synthetic samples were then generated for minority-class instances based on their local neighbourhood difficulty ratios (Equations (1)–(3)), and the WAWQI values of the newly synthesised observations were assigned by linear interpolation between the selected minority instance and its randomly chosen nearest neighbour in the feature space. Following this augmentation step, all five quality classes achieved equal representation ($n = 164$ per class), resulting in a balanced training dataset of 820 observations (Table 2). Regression models were subsequently trained on this augmented dataset using continuous WAWQI values as the target variable, and model performance was evaluated using standard regression metrics—MAE, MSE, RMSE, and R^2 —computed on the original, unaugmented test set to ensure unbiased assessment of predictive accuracy [15,57]. This two-stage approach, combining class-informed oversampling with continuous-value regression, is consistent with recent practice in environmental data modelling where severe distributional imbalance would otherwise compromise the reliability of minority-range predictions [20,21].

2.4.3. Model Training

Multiple regression techniques were trained on the preprocessed dataset. For SVR, the radial basis function kernel was used by default. PCA was optionally applied to retain 95% of variance before SVR training to investigate dimensionality reduction effects. Additionally, an XGBoost regressor was fitted directly on the standardized features to exploit its gradient boosting framework. Finally, an LSTM was implemented for comparative time-series-inspired modeling: input sequences were reshaped into three-dimensional arrays and trained over 50 epochs with the Adam optimizer minimizing MAE loss. The dataset is divided into 75% training and 25% testing.

2.4.4. Hyperparameter Optimization

To enhance model generalizability, two search strategies were employed for both SVR and XGBoost regressors. GridSearchCV explored predefined hyperparameter grids (e.g., $C = [10, 100, 1000]$, $\gamma = [0.1, 0.01, 0.001]$, $\text{kernel} \in \{\text{rbf}, \text{linear}\}$) using five-fold cross-validation and r^2 scoring. RandomizedSearchCV sampled from broader distributions (e.g., SVR's *epsilon*, XGBoost's *n_estimators*, *learning_rate*, *max_depth*) across 30 random combinations, optimizing for negative MAE. The best hyperparameters selected by each search

were refitted on the full training set before final evaluation. The hyperparameter search ranges were established based on a combination of empirical heuristics from previous environmental modeling studies and preliminary tuning phases. These ranges were designed to be sufficiently broad to capture the global optimum while ensuring the ‘best’ parameters remained within the search boundaries, thereby avoiding local minima and ensuring model stability.

2.4.5. Model Performance Validation

To ensure robustness beyond a single train–test split, ten-fold cross-validation is conducted on the optimal SVR and XGBoost configurations, providing distributions of r^2 scores across folds. Additionally, parity plots (predicted vs. actual) and LSTM loss curves (training vs. validation) were generated to visually inspect bias, variance, and convergence behavior. Such validation steps confirm that selected models maintain consistent performance and avoid overfitting, supporting their suitability for reliable WAWQI prediction. K-fold cross-validation is a standard approach for model evaluation that involves the iterative division of the dataset into K equal-sized folds. In each iteration, one fold functions as the validation group, while the remaining K-1 folds comprise the training set. Following the completion of K iterations, the average of the validation results is calculated to provide an overall performance estimate; [58] analyzed the structure of variance components specific to this procedure, while [34] measured how varying K values affect prediction bias. These studies collectively reinforce the methodological rigor of cross-validation [57]. The selection of a smaller K can result in a reduction in computational demands; however, it may concurrently lead to an augmentation of variability in performance estimates. Consequently, a balance must be struck between efficiency and reliability. In the WAWQI context, K-fold cross-validation offers a multifaceted approach to assessing model performance, mitigating the risks of overfitting, and enhancing the generalizability of the model. The present study utilizes a range of evaluation criteria, including the R^2 , the MAE, MSE and RMSE, to provide a thorough validation of prediction accuracy. To ensure that the high predictive accuracy (R^2) achieved by the models was not a result of overfitting, several preventive measures were integrated into the workflow. These included the use of a 10-fold cross-validation scheme to assess stability across data subsets, the isolation of a 25% test set for final evaluation, and the implementation of ADASYN oversampling to prevent class-based bias. Furthermore, the convergence of training and validation loss curves (Figure 6) was monitored to verify that the models retained robust generalization capabilities on unseen data.

2.5. Model Performance Metrics

The model choices are thoroughly assessed utilizing MAE, MSE, RMSE, and the R^2 . The magnitude of the mistakes can be intuitively seen from the MAE and RMSE. The R^2 can be seen as the percentage of the data variation that the model can account for and provides a proportionate explanation of the model’s predictive power [15]. The target index WAWQI in the training models, which is excluded to facilitate direct comparison between modeling approaches, is the basis of the calculation of these criteria.

The MAE measures the average magnitude of the errors between predicted values $\hat{y}_i$ and observed values y_i , without considering their direction. It is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

A lower MAE indicates that, on average, predictions are closer to the true values. Meanwhile, MSE represents the average of the squared differences between predicted and observed values:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

By squaring the errors, larger deviations are penalized more heavily. A smaller MSE signifies better predictive accuracy. The RMSE is the square root of the MSE:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Since it is expressed in the same units as the target variable, RMSE is often more interpretable than MSE. Lower RMSE values indicate more accurate predictions. The R^2 score represents the proportion of variance in the observed data y_i that is explained by the model's predictions:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

where $\bar{y}$ is the mean of the observed values. An R^2 of 1 indicates perfect prediction, while 0 means the model performs no better than simply using the mean of the observations. Negative values imply the model fits worse than a horizontal line at $\bar{y}$.

2.6. Weighted Arithmetic Water Quality Index (WAWQI)

The water quality is calculated using the arithmetic weighted approach. Based on the framework by Brown et al. (1972) [12], the WAWQI method evaluates purity levels by synthesizing data from selected water quality parameters. Based on the calculated values the water quality could be classified into five levels: Excellent (0–25), good (25–50), poor (50–75), very poor (75–100), and unsuitable (>100). Its formula is as follows [59]:

$$\text{WAWQI} = \frac{\sum_{i=1}^n W_i Q_i}{\sum_{i=1}^n W_i} \quad (8)$$

The quality rating scale (Q_i) for each parameter is calculated by using the following formula:

$$Q_i = 100 \left[\frac{V_i - V_o}{S_i - V_o} \right] \quad (9)$$

where V_i is estimated concentration of i th parameter in the analyzed water, V_o is the ideal value of this parameter in pure water ($V_o = 0$ in general), S_i is recommended standard value of i th parameter. The unit weight (W_i) for each the water quality parameter is calculated by using the following formula:

$$W_i = \frac{K}{S_i} \quad (10)$$

where K = proportionality constant and can also be calculated by using the following formula:

$$K = \frac{1}{\sum (\frac{1}{S_i})} \quad (11)$$

3. Results and Discussion

3.1. Descriptive Summary of Input Parameters

The WAWQI assessment of analyzed samples revealed that most water samples were classified as “good” (53.2%) and “excellent” (38.3%), while only a small fraction fell under

“poor” (3.9%), “very poor” (2.3%), and “unsuitable” (2.3%) categories. This distribution indicates that overall water quality in the study area is generally acceptable, with localized issues requiring attention. Furthermore, the parameter sub-indexing and weighting process often causes eclipsing and ambiguity issues that lead to misclassification of water quality [16,60,61]. Eclipsing in water quality models describes a condition in which the computed index value misrepresents the true water quality, resulting in either overestimation or underestimation. This phenomenon may occur when specific parameters exceed critical threshold values [61,62] or as a consequence of improper sub-index formulation and parameter weight allocation. Parameter sub-indexing standardizes measured values into unitless indices, whereas parameter weighting determines the relative significance of each parameter in the overall WQI calculation [60]. Despite their necessity, both processes are recognized as major sources of eclipsing and uncertainty in WQI modeling [10,16,60,61]. Consequently, water quality representation may become distorted, as a single parameter can dominate the overall index value and mask the contribution of other important parameters [10,60,63]. However, the pronounced imbalance among WAWQI classes poses a significant challenge for ML models, as algorithms tend to favor majority classes, leading to biased predictions. To address this, the ADASYN method was applied, generating synthetic samples for minority classes based on their classification difficulty to achieve a uniform distribution across all categories ($n = 164$ per class) (Table 2). This adaptive oversampling strategy is critical for improving model generalization and ensuring unbiased classification performance in subsequent predictive modeling, as recent studies confirm that ADASYN significantly enhances classifier robustness in imbalanced datasets [56].

The descriptive statistics for the 16 parameters used in the model development are summarized in Table 3. Analysis of 308 water samples revealed a wide range of physico-chemical characteristics, essential for robust model training. The water temperature ranged from 8.4 °C to 22.11 °C. Notably, the mean temperature (21.94 °C) is very close to the maximum value, suggesting a negatively skewed distribution or the influence of localized thermal sources. The pH levels remained stable and within potable limits (mean = 7.27), while EC and TDS showed significant variance, reflecting diverse mineralogical signatures across sampling sites. SO_4^{2-} and Cl^- concentrations averaged 33.34 mg/L and 19.64 mg/L, respectively. A significant anomaly was observed in F^- concentrations, where a maximum of 452 mg/L far exceeded the mean of 1.83 mg/L, indicating heavy skewness and the presence of extreme outliers. Ca^{2+} and Na^+ exhibited relatively controlled distributions. However, Mg^{2+} and K^+ displayed the highest standard errors (SEs), signaling high dispersion and variability in the source water composition. THM concentrations spanned a broad range from 0.95 to 526 µg/L (mean = 17.83 µg/L), providing a robust gradient for predictive modeling. Conversely, nitrogen species such as NH_4^+ and NO_2^- remained consistently low, suggesting minimal organic pollution in the majority of samples.

The significant skewness observed in parameters such as F^- and K^+ justifies the requirement for a non-linear modeling approach to accurately capture the underlying water quality dynamics. The utilization of SVR, XGBoost, and LSTM models is specifically intended to address the non-linear complexities and high variance identified in the water quality dataset. While traditional linear models often fail to account for extreme fluctuations in parameters like F^- and K^+ , the kernel-based mapping of SVR, the ensemble gradient boosting architecture of XGBoost, and the deep recurrent layers of LSTM provide the necessary computational flexibility to capture these underlying stochastic dynamics.

Figure 3 presents the Pearson correlation heatmap for all 19 physicochemical and microbiological parameters analysed in this study. The matrix reveals a well-defined correlation structure that provides important insights into the geochemical and operational controls governing water quality in the study area. A near-perfect positive correlation

between electrical conductivity (EC) and total dissolved solids (TDS) ($r \approx 1.00$) confirms that ionic strength is the dominant control on water mineralization within the studied samples. Strong positive correlations of EC and TDS with Ca^{2+} ($r \approx 0.73$) and SO_4^{2-} ($r \approx 0.75$) indicate that divalent ions, particularly calcium and sulfate, play a central role in governing overall water chemistry. The positive association between Ca^{2+} and SO_4^{2-} ($r \approx 0.66$) suggests geochemical controls related to the dissolution of calcium–sulfate-bearing minerals, consistent with sulfate-rich lithological formations and elevated hardness levels. In contrast, Na^+ and K^+ form a secondary correlation group ($r \approx 0.58$), reflecting their shared alkali metal characteristics, similar geochemical behavior, and potential common mineral sources. Microbiological parameters (Total Coliforms, *E. coli*, and *Enterococci*) exhibit strong inter-correlations but weak associations with chemical variables, indicating that microbial contamination follows largely independent processes and should be evaluated separately from physicochemical water quality indicators. From a modelling perspective, the strong intercorrelations among EC, TDS, Ca^{2+} , and SO_4^{2-} confirm the suitability of PCA for dimensionality reduction in this dataset, as these collinear variables are expected to project onto shared principal components, thereby reducing redundancy without substantial information loss [27,50].

Table 3. Summary statistics of input water quality parameters.

Parameter	Min	Max	Mean	SE
Temp (°C)	8.4	22.11	21.9385	0.7141
Cl_2 (mg/L)	0	7.7	1.0549	0.0602
pH	6.69	8.13	7.2652	0.0133
EC ($\mu\text{S}/\text{cm}$)	12	962	434.1299	8.6967
TDS (mg/L)	7.8	625.3	282.1844	5.6529
NO_2^- (mg/L)	0.001	0.22	0.0197	0.0015
NO_3^- (mg/L)	0.01	40.1	7.3363	0.4411
NH_4^+ (mg/L)	0	0.65	0.0667	0.0043
SO_4^{2-} (mg/L)	1	141	33.341	1.4543
Cl^- (mg/L)	3.3	34.5	19.6414	0.3198
Ca^{2+} (mg/L)	15.63	114.02	47.3683	0.9933
Mg^{2+} (mg/L)	9.85	632	34.4909	2.5824
Na^+ (mg/L)	0.638	28	3.9268	0.2227
K^+ (mg/L)	0.998	652	14.5928	3.3937
F^- (mg/L)	0	452	1.827	1.4703
THM ($\mu\text{g}/\text{L}$)	0.95	526	17.8293	2.8208

A notable observation from the correlation analysis is the relatively strong positive association between K^+ and THMs ($r \approx 0.75$). Although potassium does not directly participate in THM formation reactions, this relationship may reflect indirect associations related to water source characteristics, treatment practices, or distribution system conditions under which both parameters exhibit concurrent variation. Elevated K^+ concentrations may serve as an indicator of broader hydrochemical or operational conditions that also influence disinfection by-product formation. When interpreted in conjunction with multivariate analyses, such co-variation is likely to be captured within the same principal component in PCA, representing a combined influence of treatment- and distribution-related factors rather than a direct geochemical control. Accordingly, the observed K^+ –THMs relationship

should be interpreted with caution and viewed as an indicator of complex underlying interactions between water chemistry and disinfection processes, rather than as evidence of a direct causal linkage.

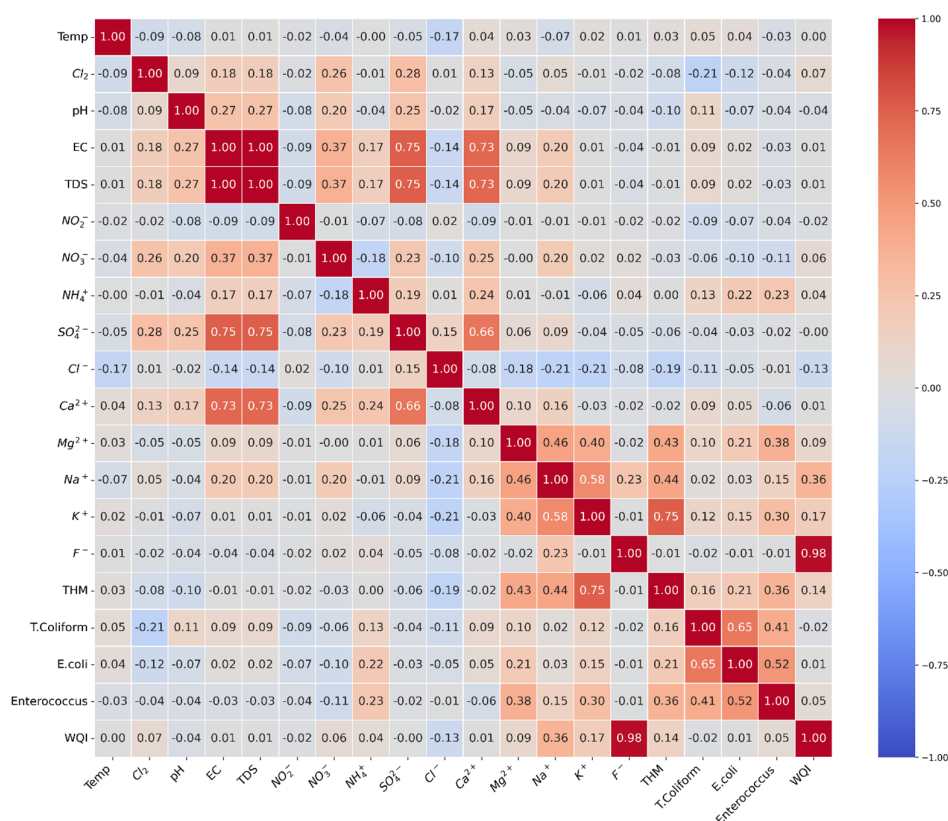


Figure 3. Pearson correlation heatmap of the 19 physicochemical and microbiological parameters measured across 308 drinking water samples. The colour scale ranges from -1 (strong negative correlation, dark blue) to $+1$ (strong positive correlation, dark red). Parameter pairs with $|r| \geq 0.70$ are considered strongly correlated. A near-perfect positive correlation is observed between EC and TDS ($r \approx 1.00$), consistent with the TDS calculation protocol employed ($\text{TDS} = k \times \text{EC}$, $k = 0.65$). Microbiological parameters (Total Coliform, *E. coli*, *Enterococci*) form a spatially distinct cluster characterised by strong mutual correlations but weak associations with physicochemical variables, indicating independent contamination pathways. The strong K^+ –THMs association ($r \approx 0.75$) is interpreted as an indirect indicator of treatment and distribution system conditions rather than a direct geochemical relationship (see Section 3.1 for full interpretation).

3.2. PCA Explained Variance Analysis

The Principal Component Analysis (PCA) was conducted to reduce the dimensionality of the dataset, and the results are summarized in Figure 4. According to the analysis, the first five principal components (PCs) have eigenvalues greater than 1, satisfying the Kaiser Criterion for component retention. As illustrated in the Screen Plot (Figure 4), a distinct ‘elbow’ point is observed at the fifth component, after which the curve begins to level off. These five components collectively account for 82.3% of the total variance in the water quality dataset. PC1, which explains 32.1% of the variance, shows high loadings for mineralization-related parameters such as EC and TDS, while subsequent components capture variance related to other physicochemical factors. By transforming the 19 original variables into these five uncorrelated latent factors, the SVR-PCA model effectively reduces multicollinearity and enhances the predictive stability of the machine learning process. By condensing the 19 original features into five uncorrelated components, the PCA-SVR model mitigated multicollinearity and enhanced predictive stability.

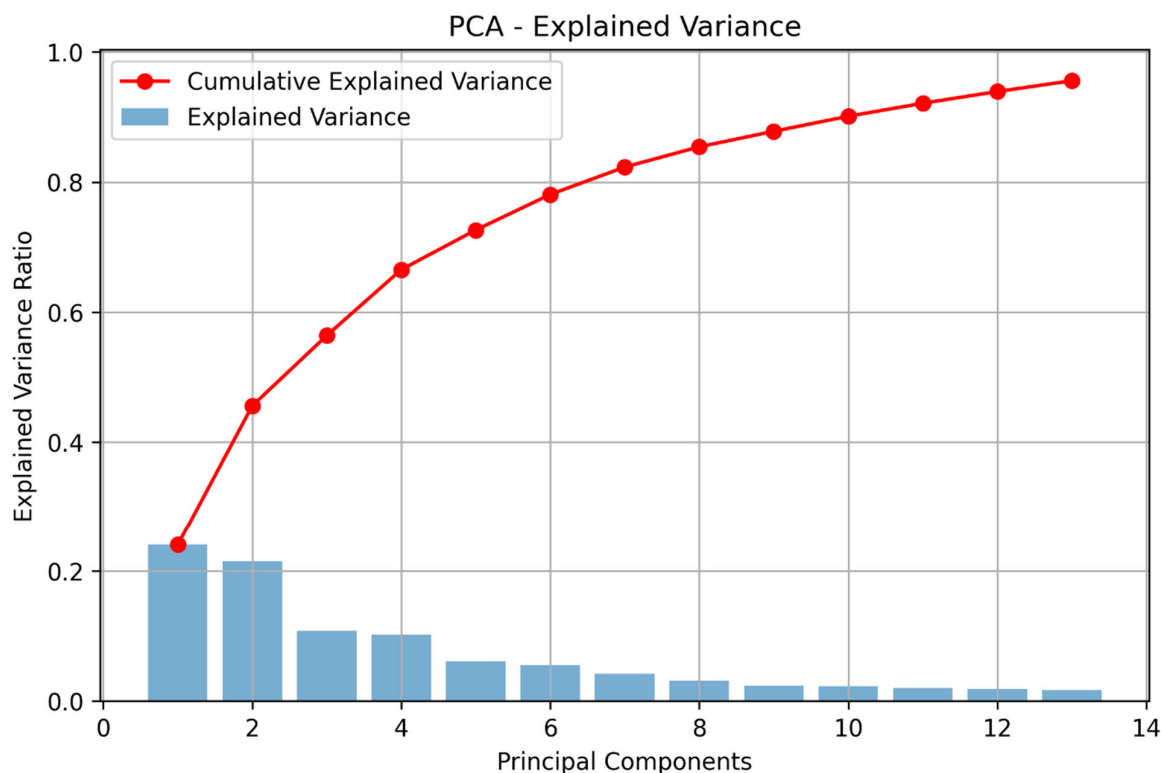


Figure 4. Cumulative Explained Variance from Principal Components Analysis (PCA) Results.

3.3. R^2 for WAWQI Prediction

The optimized LSTM model combined with Randomized Search achieved the highest predictive performance among all tested algorithms, explaining 99.9% of the variance ($R^2 = 0.999$) and yielding very low error metrics (MAE = 0.829, MSE = 1.454, RMSE = 1.206). Such performance demonstrates the capacity of deep learning to capture the high-dimensional, non-linear dynamics often found in hydrochemical data. Compared to baseline models, which exhibited moderate predictive accuracy (e.g., untuned SVR, $R^2 = 0.842$), the tuned LSTM demonstrated a substantial improvement in both accuracy and robustness. These findings suggest that sequence-based learning approaches, when coupled with adaptive hyperparameter optimization, can provide highly accurate predictions in dynamic water-quality modeling applications (Figure 5).

WAWQI, which may be defined as a rating reflecting the composite influence of different water quality parameters, is one of the most commonly used drinking water quality indices [6,59,64,65]. Conventional WQI frameworks rely on fixed parameter weights and aggregation functions that often oversimplify non-linear interactions and introduce subjectivity into water quality assessment. Indicator weight assignment is often influenced by subjective judgment, which can substantially affect the accuracy of assessment outcomes [66]. Moreover, these approaches are limited in their capacity to efficiently process large, heterogeneous datasets that vary across time and space. ML-based methods directly address these shortcomings by learning data-driven patterns without predefined assumptions, capturing complex variable interactions, and identifying influential parameters with high predictive accuracy. Their adaptability to continuous data updates further supports dynamic and reliable water quality assessment, offering a robust decision-support tool for sustainable water resource management and policy development. Therefore, ML algorithms have demonstrated strong potential for handling large datasets and generating reliable predictions to support informed decision-making [19]. The limitations of conventional WQI methodologies—including subjectivity in parameter weighting and their

inability to effectively handle large, complex datasets—highlight a clear need for robust ML-based approaches to improve the accuracy and reliability of water quality prediction. The observed reduction in performance for the XGBoost + PCA model compared to its baseline counterpart suggests that dimensionality reduction may lead to the loss of informative variance essential for tree-based ensemble learning. While PCA effectively addressed the high correlation between ionic parameters like EC and TDS, it potentially discarded critical signals within the 5% unexplained variance, which included localized extreme fluctuations in Fluoride and THMs. Consequently, for models with inherent feature selection capabilities, using the full 19-parameter feature set proved more robust than using PCA-transformed latent factors.

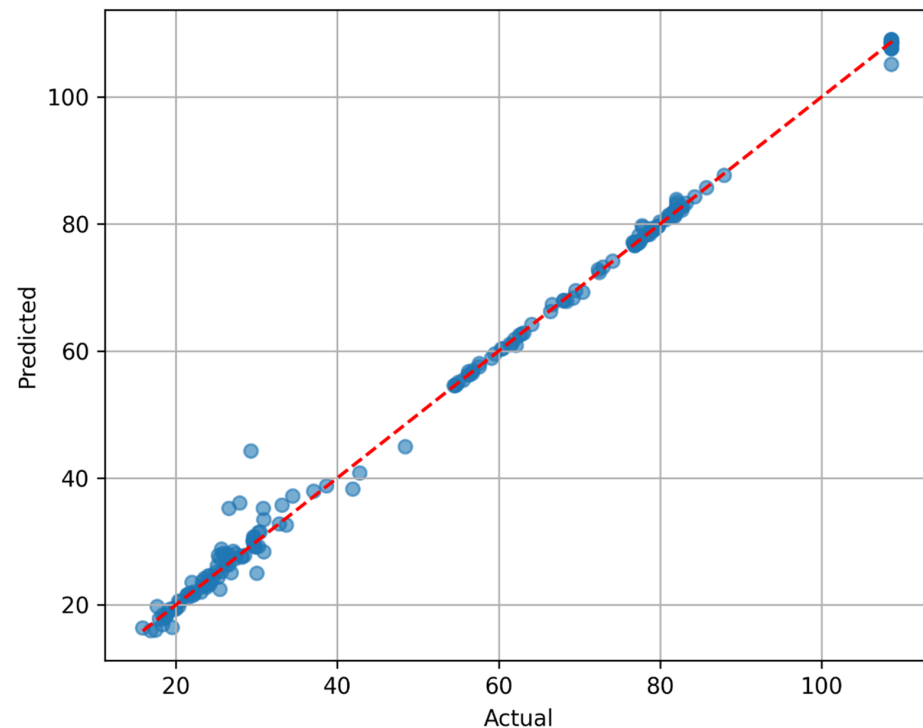


Figure 5. Comparative Performance of Different Machine Learning Algorithms Based on R^2 Values. The red dashed line represents a perfect correlation where actual and predicted values are equal. The blue dots depict the model's actual values versus their corresponding predicted values.

3.4. Key Findings on Model Performance

In evaluating predictive performance beyond the R^2 , three complementary error metrics—MAE, MSE, and RMSE—offer critical insight into the average deviations, the impact of large residuals, and the scale of prediction errors in the context of water-quality modelling.

3.4.1. Evaluation of Performance in Terms of the MAE Metric

MAE provides an intuitive measure of prediction accuracy by quantifying the average absolute deviation between predicted and observed WAWQI values, expressed in the same units as the target variable. As shown in Table 4 the baseline SVR model exhibits a high MAE of 8.699, indicating substantial prediction error. Incorporating PCA yields negligible improvement (MAE = 8.708), suggesting that dimensionality reduction alone does not significantly enhance SVR performance. In contrast, hyperparameter optimization markedly improves accuracy: SVR combined with GridSearch reduces MAE to 0.752, while RandomizedSearch achieves 0.890, representing an error reduction of over 90% compared to the untuned model. XGBoost demonstrates superior baseline performance (MAE = 0.750)

and maintains competitive accuracy after tuning, with GridSearch and RandomizedSearch yielding MAE values of 0.909 and 0.813, respectively. Although XGBoost + PCA performs poorly (MAE = 2.343), optimized configurations rival SVR in precision. Finally, the LSTM model attains an MAE of 0.981, and its tuned variant achieves 0.829, underscoring the capacity of deep learning architectures to capture temporal dependencies and deliver robust predictions without extensive feature engineering.

Table 4. Comparison of Performance Metrics for WAWQI Estimation of ML Models.

Model	R ²	MAE	MSE	RMSE
SVR	0.842	8.699	166.527	12.905
SVR + PCA	0.843	8.708	165.669	12.871
SVR + GridSearch	0.997	0.752	2.845	1.687
SVR + RandomizedSearch	0.995	0.890	5.542	2.354
XGBoost	0.998	0.750	2.024	1.423
XGBoost + PCA	0.980	2.343	21.569	4.644
XGBoost + GridSearch	0.998	0.909	2.271	1.507
XGBoost + RandomizedSearch	0.998	0.813	2.091	1.446
LSTM	0.998	0.981	2.173	1.474
LSTM + RandomizedSearch	0.999	0.829	1.454	1.206

3.4.2. Evaluation of Performance in Terms of the MSE Metric

MSE accentuates larger deviations by squaring residuals before averaging, making it highly sensitive to outliers and thus a rigorous indicator of model robustness. As shown in Table 4 the untuned SVR exhibits an extremely high MSE of **166.527**, reflecting substantial prediction discrepancies. Incorporating PCA yields only a marginal improvement (MSE = 165.669), indicating that dimensionality reduction alone does not significantly enhance SVR performance. In contrast, hyperparameter optimization produces dramatic gains: SVR combined with GridSearch reduces MSE to 2.845, while RandomizedSearch achieves 5.542, representing a reduction of more than 95% compared to the baseline. XGBoost demonstrates superior baseline performance (MSE = 2.024) and maintains competitive accuracy after tuning, with GridSearch and RandomizedSearch yielding MSE values of 2.271 and 2.091, respectively. Although XGBoost + PCA performs poorly (MSE = 21.569), optimized configurations rival SVR in precision. Finally, the LSTM model attains an MSE of 2.173, and its tuned variant achieves 1.454, underscoring the capacity of deep learning architectures to capture temporal dependencies and deliver robust predictions without extensive feature engineering.

3.4.3. Evaluation of Performance in Terms of the RMSE Metric

RMSE converts the squared deviations of MSE back into the original units of the target variable, providing an interpretable measure of prediction accuracy that complements MAE and MSE. As shown in Table 4, the untuned SVR exhibits a very high RMSE of 12.905, reflecting substantial deviation from observed WAWQI values. Incorporating PCA yields only a negligible improvement (RMSE = 12.871), indicating that dimensionality reduction alone does not significantly enhance SVR performance. In contrast, hyperparameter optimization produces dramatic gains: SVR combined with GridSearch reduces RMSE to 1.687, while RandomizedSearch achieves 2.354, representing a reduction of more than 80% compared to the baseline. XGBoost demonstrates superior baseline performance (RMSE = 1.423) and maintains competitive accuracy after tuning, with GridSearch and RandomizedSearch

yielding RMSE values of 1.507 and 1.446, respectively. Although XGBoost + PCA performs poorly ($\text{RMSE} = 4.644$), optimized configurations rival SVR in precision. Finally, the LSTM model attains an RMSE of 1.474, and its tuned variant achieves 1.206, underscoring the capacity of deep learning architectures to capture temporal dependencies and deliver robust predictions without extensive feature engineering.

3.5. Evaluation of LSTM + RandomizedSearch Model Convergence

Figure 6 illustrates the training dynamics of the optimized LSTM model, which achieved the highest predictive performance among all tested algorithms. The loss curve, plotting training and validation loss across successive epochs, demonstrates a consistent and gradual decline for both metrics, indicating effective learning and stable convergence. During the initial epochs, a sharp reduction in loss is observed, reflecting rapid adjustment of network weights during early optimization. As training progresses, the loss values plateau, suggesting that the model approaches its minimum error configuration. Importantly, the validation loss closely tracks the training loss throughout the process, with no significant divergence, providing strong evidence against overfitting.

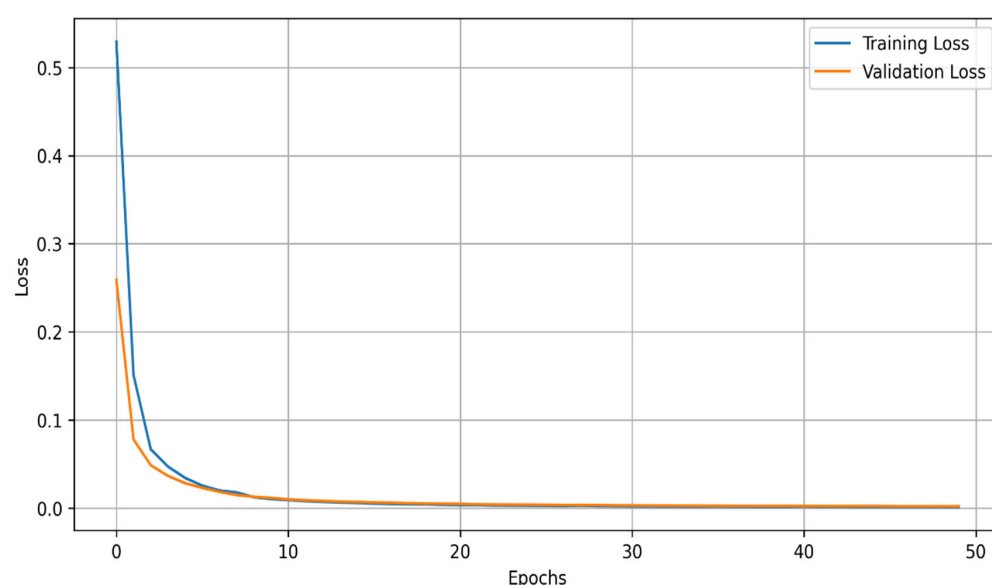


Figure 6. Training and Validation Loss Curves of the Optimized LSTM Model. The LSTM Model Loss metric is indicative of the loss and validation loss of the LSTM model during the training process. The X-axis of the graph corresponds to the “Epochs” (training cycles), while the Y-axis corresponds to the “Loss” values.

The final epoch loss values for both training and validation remain low and closely aligned, reinforcing the model’s generalization capability on unseen data. This convergence behavior confirms that the LSTM architecture, when combined with adaptive hyperparameter tuning (e.g., optimizer selection, batch size, and epoch count), is highly effective in capturing temporal dependencies and non-linear relationships inherent in WAWQI datasets. These characteristics explain why the tuned LSTM achieved superior accuracy and minimal error metrics compared to other machine learning approaches.

The analytical reliability of the dataset underpins the validity of the predictive models developed in this study. All physicochemical and microbiological parameters were measured in accordance with internationally recognised protocols—specifically ITASHY regulations and USEPA Method 200.7—at the accredited Şanlıurfa Water and Sewerage Administration Drinking Water Laboratory [41,44]. In situ measurements of pH, EC, free chlorine, and temperature were conducted using factory-calibrated portable instruments,

with recalibration performed prior to each sampling campaign. The broad parameter ranges and pronounced skewness documented in Table 3—particularly for F^- , K^+ , Mg^{2+} , and THMs—are therefore attributable to genuine spatial and seasonal heterogeneity across the 13 districts rather than to analytical artefacts, a conclusion supported by the internal consistency of inter-parameter correlations observed in Figure 3. Nonetheless, in line with recommendations by Uddin et al. (2022) [16] and Reviewer guidance, the absence of duplicate sample analyses and certified reference material recovery data constitutes a limitation of the present dataset, which should be addressed in future monitoring campaigns to further strengthen the analytical underpinning of data-driven water quality models.

For the purpose of data collection, 308 water samples were obtained from 77 sampling locations across 13 districts of Şanlıurfa between January and September 2018 at regular intervals, covering all four seasons to capture temporal variability in water quality. Sampling and analyses were conducted in accordance with standardized procedures to ensure data reliability and consistency. In situ measurements of pH, EC, Cl^- , and temperature were performed using calibrated portable instruments. Laboratory analyses of physicochemical parameters were carried out at the Şanlıurfa Water and Sewerage Administration Drinking Water Laboratory using spectrophotometric techniques and ICP-OES, following ITASHY guidelines and the USEPA 200.7 method.

The dataset was subsequently classified using a WAWQI approach, categorizing samples into five quality classes: Excellent, Good, Poor, Very Poor, and Unsuitable. The majority of samples were classified as Good (53.2%) or Excellent (38.3%), while only a limited proportion fell into the Poor (3.9%), Very Poor (2.3%), and Unsuitable (2.3%) categories (Table 2). This distribution indicates that drinking water quality in the study area is generally acceptable, although localized degradation may exist. To enhance the predictive robustness of the ML models and mitigate class imbalance effects, adaptive oversampling techniques were applied to balance WAWQI classes, thereby ensuring equitable representation and improving model generalization.

This study aimed to identify the most suitable ML-based approach for WAWQI estimation through a comprehensive comparison of different algorithms. Previous studies have demonstrated the effectiveness of ML models in water quality assessment and WQI prediction [67–69], with recent research emphasizing the benefits of combining classical statistical techniques with advanced learning frameworks [70–72]. In addition, hybrid ML approaches have been successfully applied in groundwater potential mapping and environmental modeling [73–75]. In this context, the present study contributes by evaluating multiple ML configurations for predicting the WAWQI.

Model performance was assessed using standard regression metrics, including MAE, MSE, RMSE, and R^2 (Equations (4)–(7)), allowing for an objective comparison of predictive accuracy. Among the tested models, SVR exhibited strong performance, particularly when coupled with hyperparameter optimization. The SVR model optimized via GridSearch achieved an R^2 value of 0.997, outperforming its RandomizedSearch counterpart ($R^2 = 0.995$) and highlighting the importance of systematic parameter tuning. XGBoost demonstrated consistently high predictive accuracy ($R^2 = 0.998$), although a slight reduction in performance was observed following PCA integration ($R^2 = 0.980$), suggesting that dimensionality reduction may lead to the loss of informative variance in this dataset.

The LSTM architecture achieved competitive predictive performance ($R^2 = 0.998$), which further improved following optimization with RandomizedSearch ($R^2 = 0.999$). This improvement indicates that sequence-based deep learning models can effectively capture complex non-linear relationships and sequential patterns embedded in water-quality data. Compared to SVR + GridSearch ($R^2 = 0.997$, MAE = 0.752, RMSE = 1.687), the optimized LSTM model yielded slightly lower error metrics (MAE = 0.829, MSE = 1.454,

RMSE = 1.206), reflecting enhanced robustness in modeling dynamic variations. The convergence behavior of the training and validation loss functions further suggests stable learning without pronounced overfitting. Stable convergence of both training and validation loss functions confirms the model's robust generalization without signs of overfitting.

3.6. Comparative Analysis of Model Properties, Advantages, and Limitations

The comparative analysis highlighted clear differences among the models. LSTM achieved the highest predictive accuracy ($R^2 = 0.999$) due to its ability to capture non-linear temporal dependencies in water quality data. XGBoost showed strong computational efficiency and robust performance with raw features, benefiting from internal regularization (L_1/L_2) that mitigates multicollinearity without requiring PCA. SVR also demonstrated strong generalization capability in high-dimensional spaces when optimized via Grid Search ($R^2 = 0.997$). However, some limitations were observed. LSTM performance was sensitive to the structure and balance of temporal data and required ADASYN oversampling for the 308-sample dataset. In addition, PCA did not consistently improve model performance and, in some cases, reduced accuracy by removing informative variance, indicating that original features may be preferable for high-gradient environmental datasets.

The superior performance of the LSTM + Randomised Search model ($R^2 = 0.999$, RMSE = 1.206) is consistent with recent comparable studies. Han et al. (2025) [19] working on urban river systems in China, reported LSTM R^2 values above 0.996, confirming that recurrent architectures are particularly effective where water quality data exhibit seasonal temporal dependencies. Similarly, Elmotawakkil et al. (2025) [20], using drinking water monitoring data from Morocco, identified LSTM as the best-performing model ($R^2 = 0.9999$), with XGBoost achieving classification accuracy of 99.07–99.99%—a pattern in strong agreement with the present study. In contrast, Nishat et al. (2025) [35], comparing fourteen ML algorithms for WQI prediction in Dhaka, Bangladesh, reported a best R^2 of 0.971 using ANN, whilst Talukdar et al. (2024) [25], optimising ML models for Lake Loktak, India, found RF superior at $R^2 = 0.97$; both studies further confirm that ensemble and deep learning architectures consistently outperform kernel-based approaches in complex multi-parameter datasets. The relatively higher absolute RMSE observed in the present study compared to Han et al. (2025) [19] reflects differences in target variable scale rather than reduced model accuracy [15,17]. Whilst deep learning demonstrated superior predictive capability, hyperparameter-optimised XGBoost remains a computationally efficient and interpretable alternative where model transparency is prioritised [21,36].

To contextualise the predictive performance reported in the present study, Table 5 provides a structured comparison of the best-performing model—LSTM + Randomised Search ($R^2 = 0.999$, MAE = 0.829, RMSE = 1.206)—against findings from recent ML-based WQI prediction studies conducted in diverse geographic and climatic settings. The comparison encompasses studies published between 2022 and 2025 that employed regression-based evaluation metrics (R^2 , RMSE, MAE) consistent with those used herein, thereby enabling a methodologically coherent benchmarking exercise.

The results presented in Table 4 demonstrate that the LSTM + Randomised Search model developed in the present study achieves performance metrics that are broadly consistent with—and in several cases superior to—those reported in the recent literature. Han et al. (2025) [19], working on WQI prediction in urban Chinese river systems using LSTM combined with climate indicators, reported R^2 values consistently above 0.996 and RMSE as low as 0.061 across training, testing, and validation datasets, confirming LSTM's capacity to model complex temporal dependencies in water quality data. Similarly, Elmotawakkil et al. (2025) [20], evaluating multiple ML algorithms for WQI prediction and classification using monitoring station data from Morocco, identified LSTM as the top-performing model

with an R^2 of 0.9999, whilst XGBoost demonstrated classification accuracy in the range of 99.07–99.99%; these findings are in strong agreement with the present study's results for both LSTM and XGBoost configurations. In contrast, studies employing non-deep-learning approaches have reported comparatively lower performance. Nishat et al. (2025) [35], applying fourteen ML models to predict WQI in the heavily polluted rivers of Dhaka, Bangladesh, reported a best R^2 of 0.971 and RMSE of 2.34 using ANN, whilst SVR performed poorly—a pattern consistent with the baseline SVR performance observed in the present study prior to hyperparameter optimisation. Talukdar et al. (2024) [25], optimising and interpreting ML models for WQI management in Lake Loktak, India, found RF to be the superior model with an R^2 of 0.97, further supporting the observation that ensemble and deep learning architectures consistently outperform kernel-based methods in complex, multi-parameter water quality datasets.

Table 5. Comparison of model performance metrics for WQI/WAWQI prediction across recent studies.

Study	Location	Best Model	R^2	RMSE	MAE	Target
Han et al. (2025) [19]	Urban rivers, China	LSTM	>0.996	0.061–0.081	—	WQI
Elmotawakkil et al. (2025) [20]	Morocco	LSTM	0.9999	—	—	WQI
Nishat et al. (2025) [35]	Dhaka rivers, Bangladesh	ANN	0.971	2.34	1.24	WQI
Mo et al. (2024) [14]	Coastal city, China	Ensemble ML	—	—	—	WQI
Talukdar et al. (2024) [25]	Lake Loktak, India	RF	0.97	—	—	WQI
This study	Şanlıurfa, Türkiye	LSTM + RS	0.999	1.206	0.829	WAWQI

Collectively, these comparisons indicate that the predictive accuracy achieved by the LSTM + Randomised Search framework in the present study is not only methodologically reproducible but also geographically transferable across diverse hydrological contexts—including urban Chinese rivers, Moroccan monitoring networks, Indian lacustrine systems, and South Asian fluvial environments—provided that adequate temporal structure and sufficient sample representativeness are maintained in the training dataset. The relatively higher RMSE observed in the present study compared to Han et al. (2025) [19] is attributable to the differences in the target variable scale; WAWQI values in the study area span a broader numerical range than the normalised WQI scores used in that study, rendering direct RMSE comparisons between absolute values less informative than R^2 and MAE metrics [16,17].

Overall, the findings align with recent studies reporting the effectiveness of LSTM-based models in environmental and water-quality forecasting applications [13,76–78]. However, while deep learning approaches demonstrated superior predictive capability in this study, their performance should be interpreted within the context of data structure, temporal resolution, and model complexity. Consequently, the results support the use of LSTM models as a complementary tool alongside conventional ML methods, particularly in applications involving temporally structured water-quality datasets where non-linear interactions play a dominant role.

The comparative analysis revealed distinct strengths and limitations among the models. The LSTM model achieved the highest predictive accuracy ($R^2 = 0.999$) due to its gated recurrent structure, which enables the capture of non-linear temporal dependencies in water quality data. XGBoost demonstrated high computational efficiency and robust performance when trained on the original feature set, benefiting from internal regularization (L_1/L_2) that mitigates multicollinearity without requiring PCA. The SVR model also showed strong

generalization capability in high-dimensional spaces, particularly when optimized via Grid Search ($R^2 = 0.997$). However, several limitations were identified. The performance of the LSTM model was sensitive to the structure and balance of the temporal dataset and therefore required ADASYN oversampling to address class imbalance in the 308-sample dataset. Additionally, dimensionality reduction through PCA did not consistently improve predictive performance and, in some cases, reduced accuracy by discarding informative variance. This suggests that, for high-gradient environmental datasets, preserving the original feature space may be preferable to relying on latent components.

4. Conclusions

The primary aim of this study was to predict the WAWQI using ten advanced machine learning algorithms, including SVR, SVR + PCA, SVR + Grid Search, SVR + Randomized Search, XGBoost, XGBoost + PCA, XGBoost + Grid Search, XGBoost + Randomized Search, LSTM, and LSTM + Randomized Search. A comprehensive comparative evaluation was conducted to assess the predictive performance of these models based on standard error metrics and goodness-of-fit measures. Key findings are summarized as follows:

1. Among all tested algorithms, the LSTM model optimized with Randomized Search achieved the highest predictive performance. Its ability to capture temporal dependencies and complex non-linear relationships within the dataset underscores the superiority of deep learning architectures for dynamic water quality forecasting.
2. Hyperparameter optimization significantly improved the accuracy of SVR and XGBoost models compared to their baseline versions. SVR combined with Grid Search and Randomized Search demonstrated strong performance, while XGBoost maintained high accuracy after tuning.
3. The integration of PCA did not consistently enhance predictive capability, indicating that dimensionality reduction alone is insufficient for improving model performance in this context.
4. Both SVR-based models with hyperparameter tuning and LSTM offer reliable solutions for WAWQI estimation, with SVR being suitable for structured datasets and LSTM recommended for time-dependent forecasting tasks.
5. Whilst the integrated modelling framework incorporates 19 physicochemical and microbiological input variables, feature importance analysis based on XGBoost coefficients and PCA loadings indicates that EC, TDS, Ca^{2+} , SO_4^{2-} , and K^+ collectively account for the dominant variance in WAWQI outcomes. In practice, a reduced monitoring protocol targeting these five key physicochemical parameters—supplemented by periodic microbiological screening for Total Coliform, *E. coli*, and *Enterococci*—would substantially reduce operational costs whilst preserving the essential predictive information required for reliable WAWQI estimation. Dissolved inorganic nitrogen species—including NO_3^- , NO_2^- , and NH_4^+ —were incorporated as input variables and, although their mean concentrations remained low across the monitoring period, their inclusion is essential for detecting localised agricultural contamination signals, particularly given the intensive irrigation practices within the GAP region [3,40].

The findings emphasize the potential of machine learning techniques as powerful tools for water quality monitoring and management. These models can support decision-making processes by providing accurate predictions, enabling proactive strategies for pollution control, resource optimization, and public health protection in semi-arid regions. Furthermore, the demonstrated superiority of deep learning approaches suggests that their integration into water management frameworks can significantly enhance predictive precision and operational efficiency.

The geographical focus of this study on Şanlıurfa Province constitutes both a deliberate methodological choice—enabling a spatially and seasonally consistent dataset to be compiled under uniform regulatory and climatic conditions—and an acknowledged limitation with respect to broader generalisability. Semi-arid regions with comparable hydrogeological characteristics, including parts of the Middle East, North Africa, and Central Asia, share similar patterns of elevated mineralisation, seasonal hydrological variability, and intensive agricultural water use, suggesting that the proposed modelling framework may be transferable to analogous settings [28,76]. Future research should validate the framework across multiple semi-arid basins and incorporate longer monitoring periods to assess inter-annual variability and the potential influence of climate-driven shifts in water quality dynamics.

In addition, the incorporation of duplicate sample analyses and certified reference material recovery data in future studies would strengthen the analytical validation of input datasets and enhance the reproducibility of data-driven water quality modelling outcomes [16].

Author Contributions: Conceptualization, M.İ.Y.; methodology, H.Y. and İ.Ö.; software, İ.Ö.; validation, H.Y. and İ.Ö.; formal analysis, H.Y. and İ.Ö.; investigation, M.İ.Y.; resources, İ.B.; data curation, İ.B.; writing—original draft preparation, B.Ç., H.Y. and İ.Ö.; writing—review and editing, B.Ç., H.Y., İ.Ö. and M.İ.Y.; visualization, H.Y. and İ.Ö.; supervision, M.İ.Y., project administration, M.İ.Y. All authors have read and agreed to the published version of the manuscript.

Funding: In accordance with the support given to doctoral thesis, this study was funded by Harran University Scientific Research Project Coordination Office (HUBAP Project no. 25103).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their insightful comments and valuable suggestions, which significantly helped to improve the quality of this manuscript.

Conflicts of Interest: The authors declared no conflicts of interest.

References

1. Han, X.; Tang, F.; Liu, A.-L. Drinking Water Quality Evaluation in Supply Systems in Wuhan, China: Application of Entropy Weight Water Quality Index and Multivariate Statistical Analysis. *Environ. Sci. Pollut. Res.* **2023**, *31*, 280–292. [\[CrossRef\]](#)
2. İlhan, N.; Demir Yetiş, A.; Yeşilnacar, M.İ.; Atasoy, A.D.S. Predictive Modelling and Seasonal Analysis of Water Quality Indicators: Three Different Basins of Şanlıurfa, Turkey. *Environ. Dev. Sustain.* **2022**, *24*, 3258–3292. [\[CrossRef\]](#)
3. Saeed, A.; Alsini, A.; Amin, D. Water Quality Multivariate Forecasting Using Deep Learning in a West Australian Estuary. *Environ. Model. Softw.* **2024**, *171*, 105884. [\[CrossRef\]](#)
4. Wieczorek, K.; Turek, A.; Szczesio, M.; Wolf, W.M. A Holistic Approach to the Spatio-Temporal Variability Investigation of the Main River Water Quality—The Importance of Tributaries. *Sci. Total Environ.* **2024**, *906*, 167588. [\[CrossRef\]](#)
5. Kothari, V.; Vij, S.; Sharma, S.; Gupta, N. Correlation of Various Water Quality Parameters and Water Quality Index of Districts of Uttarakhand. *Environ. Sustain. Indic.* **2021**, *9*, 100093. [\[CrossRef\]](#)
6. Mukate, S.; Wagh, V.; Panaskar, D.; Jacobs, J.A.; Sawant, A. Development of New Integrated Water Quality Index (IWQI) Model to Evaluate the Drinking Suitability of Water. *Ecol. Indic.* **2019**, *101*, 348–354. [\[CrossRef\]](#)
7. Njuguna, S.M.; Onyango, J.A.; Githaiga, K.B.; Gituru, R.W.; Yan, X. Application of Multivariate Statistical Analysis and Water Quality Index in Health Risk Assessment by Domestic Use of River Water. Case Study of Tana River in Kenya. *Process Saf. Environ. Prot.* **2020**, *133*, 149–158. [\[CrossRef\]](#)
8. Babuji, P.; Thirumalaisamy, S.; Duraisamy, K.; Periyasamy, G. Human Health Risks Due to Exposure to Water Pollution: A Review. *Water* **2023**, *15*, 2532. [\[CrossRef\]](#)
9. Milijašević Joksimović, D.; Jakovljević, D.; Jojić Glavonjić, T. Evaluation of Water Quality in the Tamiš River in Serbia Using the Water Pollution Index: Key Pollutants and Their Sources. *Water* **2025**, *17*, 1024. [\[CrossRef\]](#)

10. Uddin, M.G.; Nash, S.; Olbert, A.I. A Review of Water Quality Index Models and Their Use for Assessing Surface Water Quality. *Ecol. Indic.* **2021**, *122*, 107218. [\[CrossRef\]](#)
11. Horton, R.K. An Index Number System for Rating Water Quality. *J. Water Pollut. Control Fed.* **1965**, *37*, 300–306.
12. Brown, R.M.; McClelland, N.I.; Deininger, R.A.; O'Connor, M.F. A Water Quality Index—Crashing the Psychological Barrier. In Proceedings of the Indicators of Environmental Quality: Proceedings of a Symposium Held During the AAAS Meeting, Philadelphia, PA, USA, 26–31 December 1971; Springer: Berlin/Heidelberg, Germany, 1972; pp. 173–182.
13. Rana, R.; Kalia, A.; Boora, A.; Alfaisal, F.M.; Alharbi, R.S.; Berwal, P.; Alam, S.; Khan, M.A.; Qamar, O. Artificial Intelligence for Surface Water Quality Evaluation, Monitoring and Assessment. *Water* **2023**, *15*, 3919. [\[CrossRef\]](#)
14. Mo, Y.; Xu, J.; Liu, C.; Wu, J.; Chen, D. Assessment and Prediction of Water Quality Index (WQI) by Seasonal Key Water Parameters in a Coastal City: Application of Machine Learning Models. *Environ. Monit. Assess.* **2024**, *196*, 1008. [\[CrossRef\]](#)
15. Uddin, M.G.; Rahman, A.; Rosa Taghikhah, F.; Olbert, A.I. Data-Driven Evolution of Water Quality Models: An in-Depth Investigation of Innovative Outlier Detection Approaches—A Case Study of Irish Water Quality Index (IEWQI) Model. *Water Res.* **2024**, *255*, 121499. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Uddin, M.G.; Nash, S.; Mahammad Diganta, M.T.; Rahman, A.; Olbert, A.I. Robust Machine Learning Algorithms for Predicting Coastal Water Quality Index. *J. Environ. Manag.* **2022**, *321*, 115923. [\[CrossRef\]](#)
17. Gupta, S.; Gupta, S.K. A Critical Review on Water Quality Index Tool: Genesis, Evolution and Future Directions. *Ecol. Inform.* **2021**, *63*, 101299. [\[CrossRef\]](#)
18. Chellaiah, C.; Anbalagan, S.; Swaminathan, D.; Chowdhury, S.; Kadhila, T.; Shopati, A.K.; Shangdiar, S.; Sharma, B.; Amesho, K.T.T. Integrating Deep Learning Techniques for Effective River Water Quality Monitoring and Management. *J. Environ. Manag.* **2024**, *370*, 122477. [\[CrossRef\]](#)
19. Han, Z.; Zhang, S.; He, L.; Zhu, B. Predicting and Investigating Water Quality Index by Robust Machine Learning Methods. *J. Environ. Manag.* **2025**, *381*, 125156. [\[CrossRef\]](#)
20. Elmotawakkil, A.; Enneya, N.; Bhagat, S.K.; Ouda, M.M.; Kumar, V. Advanced Machine Learning Models for Robust Prediction of Water Quality Index and Classification. *J. Hydroinform.* **2025**, *27*, 299–319. [\[CrossRef\]](#)
21. Xu, J.; Mo, Y.; Zhu, S.; Wu, J.; Jin, G.; Wang, Y.-G.; Ji, Q.; Li, L. Assessing and Predicting Water Quality Index with Key Water Parameters by Machine Learning Models in Coastal Cities, China. *Heliyon* **2024**, *10*, e33695. [\[CrossRef\]](#)
22. Hussein, E.E.; Jat Baloch, M.Y.; Nigar, A.; Abualkhair, H.F.; Aldawood, F.K.; Tageldin, E. Machine Learning Algorithms for Predicting the Water Quality Index. *Water* **2023**, *15*, 3540. [\[CrossRef\]](#)
23. Derdour, A.; Jodar-Abellan, A.; Pardo, M.Á.; Ghoneim, S.S.M.; Hussein, E.E. Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water* **2022**, *14*, 2801. [\[CrossRef\]](#)
24. Singh, S.; Das, A.; Sharma, P. Predictive Modeling of Water Quality Index (WQI) Classes in Indian Rivers: Insights from the Application of Multiple Machine Learning (ML) Models on a Decennial Dataset. *Stoch. Environ. Res. Risk Assess.* **2024**, *38*, 3221–3238. [\[CrossRef\]](#)
25. Talukdar, S.; Shahfahad; Bera, S.; Naikoo, M.W.; Ramana, G.V.; Mallik, S.; Kumar, P.A.; Rahman, A. Optimisation and Interpretation of Machine and Deep Learning Models for Improved Water Quality Management in Lake Loktak. *J. Environ. Manag.* **2024**, *351*, 119866. [\[CrossRef\]](#)
26. Mahanty, B.; Lhamo, P.; Sahoo, N.K. Inconsistency of PCA-Based Water Quality Index—Does It Reflect the Quality? *Sci. Total Environ.* **2023**, *866*, 161353. [\[CrossRef\]](#)
27. Yenigun, I.; Bilgili, A.V.; Yesilnacar, M.I.; Yalcin, H. Seasonal and Spatial Variations in Water Quality of Deep Aquifer in the Harran Plain, GAP Project, Southeastern Anatolia, Turkey. *Environ. Earth Sci.* **2021**, *80*, 568. [\[CrossRef\]](#)
28. Noori, R.; Sabahi, M.S.; Karbassi, A.R.; Baghvand, A.; Taati Zadeh, H. Multivariate Statistical Analysis of Surface Water Quality Based on Correlations and Variations in the Data Set. *Desalination* **2010**, *260*, 129–136. [\[CrossRef\]](#)
29. Gupta, S.; Nayek, S.; Chakraborty, D. Hydrochemical Evaluation of Rangit River, Sikkim, India: Using Water Quality Index and Multivariate Statistics. *Environ. Earth Sci.* **2016**, *75*, 567. [\[CrossRef\]](#)
30. Jahin, H.S.; Abuzaid, A.S.; Abdellatif, A.D. Using Multivariate Analysis to Develop Irrigation Water Quality Index for Surface Water in Kafr El-Sheikh Governorate, Egypt. *Environ. Technol. Innov.* **2020**, *17*, 100532. [\[CrossRef\]](#)
31. Ma, Z.; Li, H.; Ye, Z.; Wen, J.; Hu, Y.; Liu, Y. Application of Modified Water Quality Index (WQI) in the Assessment of Coastal Water Quality in Main Aquaculture Areas of Dalian, China. *Mar. Pollut. Bull.* **2020**, *157*, 111285. [\[CrossRef\]](#)
32. Hamma, B.; Alodah, A.; Bouaicha, F.; Bekkouche, M.F.; Barkat, A.; Hussein, E.E. Hydrochemical Assessment of Groundwater Using Multivariate Statistical Methods and Water Quality Indices (WQIs). *Appl. Water Sci.* **2024**, *14*, 33. [\[CrossRef\]](#)
33. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A Review of the Application of Machine Learning in Water Quality Evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Fushiki, T. Estimation of Prediction Error by Using K-Fold Cross-Validation. *Stat. Comput.* **2009**, *21*, 137–146. [\[CrossRef\]](#)

35. Nishat, M.H.; Khan, M.H.R.B.; Ahmed, T.; Hossain, S.N.; Ahsan, A.; El-Sergany, M.M.; Shafiquzzaman, M.; Imteaz, M.A.; Alresheedi, M.T. Comparative Analysis of Machine Learning Models for Predicting Water Quality Index in Dhaka's Rivers of Bangladesh. *Environ. Sci. Eur.* **2025**, *37*, 31. [CrossRef]
36. Bai, Y.; Peng, M.; Wang, M. A River Water Quality Prediction Method Based on Dual Signal Decomposition and Deep Learning. *Water* **2024**, *16*, 3099. [CrossRef]
37. MGM. Turkish State Meteorological Service. Climate Data. 2023. Available online: <https://www.mgm.gov.tr/veridegerlendirme/il-ve-ilceler-istatistik.aspx> (accessed on 20 July 2025).
38. Ay, M.; Sakar, E.; Ercisli, S. The Effect of Harvest Period on Total Phenolic Content, Phenolic Compounds and Peroxide Value in Oil of Four Olive Cultivars Grown in Şanlıurfa. *Appl. Fruit Sci.* **2025**, *67*, 92. [CrossRef]
39. GAP RDA. Southeastern Anatolia Project Regional Development Administration. Annual Report. 2022. Available online: <https://www.gap.gov.tr/e-kutuphane-detay/2022-yili-faaliyet-raporu/> (accessed on 20 July 2025).
40. Bayhan, İ. Şanlıurfa İli İçme Kullanma Sularının Fiziksel, Kimyasal ve Mikrobiyolojik Kalitesinin Araştırılması. Ph.D. Thesis, Harran Üniversitesi, Şanlıurfa, Türkiye, 2021.
41. Bayhan, İ.; Yazıcı-Karabulut, B.; Yesilnacar, M.; Karabulut, A. Investigation of Microbiological Quality of Drinking and Usage Water in a Fast-Growing Metropolitan City. *New Horiz. Health Sci.* **2022**, *1*, 31–69.
42. ISO 9308-1:2014(En); Water Quality—Enumeration of Escherichia Coli and Coliform Bacteria—Part 1: Membrane Filtration Method for Waters with Low Bacterial Background Flora. International Organization for Standardization (ISO): Geneva, Switzerland, 2014.
43. ISO 7899-2:2000(En); Water Quality—Detection and Enumeration of Intestinal Enterococci—Part 2: Membrane Filtration Method. International Organization for Standardization (ISO): Geneva, Switzerland, 2000.
44. Bayhan, İ.; Yesilnacar, M.İ.; Demir Yetiş, A.; Tutkun, E. An Evaluation of Drinking-Usage Water Quality in Terms of Environmental Health: A Case Study of Siverek (Şanlıurfa), Turkey. *Turk. Bull. Hyg. Exp. Biol.* **2020**, *77*, 107–120. [CrossRef]
45. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
46. Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines*; Apress: New York, NY, USA, 2015; pp. 67–80.
47. Li, H.; Tan, Y.; Zeng, D.; Su, D.; Qiao, S. Attitude-Predictive Control of Large-Diameter Shield Tunneling: PCA-SVR Machine Learning Algorithm Application in a Case Study of the Zhuhai Xingye Express Tunnel. *Appl. Sci.* **2025**, *15*, 1880. [CrossRef]
48. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2002.
49. Carniel, T.; Gastaud, C.; Dalle, J.-M. The Temporal Evolution of Venture Investment Strategies in Sector Space. *arXiv* **2019**, arXiv:1906.01980. [CrossRef]
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
52. Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative Evaluation of Machine Learning Models for Groundwater Quality Assessment. *Environ. Monit. Assess.* **2020**, *192*, 776. [CrossRef] [PubMed]
53. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
54. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
55. Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2222–2232. [CrossRef]
56. Park, H.; Kim, H. AR-ADASYN: Angle Radius-Adaptive Synthetic Data Generation Approach for Imbalanced Learning. *Stat. Comput.* **2024**, *34*, 166. [CrossRef]
57. Wong, T.-T.; Yeh, P.-Y. Reliable Accuracy Estimates from K-Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1586–1594. [CrossRef]
58. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [CrossRef]
59. Tokath, C. Application of Water Quality Index for Drinking Purposes in Dam Lakes: A Case Study of Thrace Region. *Sigma J. Eng. Nat. Sci.* **2020**, *38*, 393–402.
60. Syeed, M.M.M.; Hossain, M.S.; Karim, M.R.; Uddin, M.F.; Hasan, M.; Khan, R.H. Surface Water Quality Profiling Using the Water Quality Index, Pollution Index and Statistical Methods: A Critical Review. *Environ. Sustain. Indic.* **2023**, *18*, 100247. [CrossRef]
61. Uddin, M.G.; Nash, S.; Rahman, A.; Olbert, A.I. A Comprehensive Method for Improvement of Water Quality Index (WQI) Models for Coastal Water Quality Assessment. *Water Res.* **2022**, *219*, 118532. [CrossRef]
62. Ding, F.; Zhang, W.; Cao, S.; Hao, S.; Chen, L.; Xie, X.; Li, W.; Jiang, M. Optimization of Water Quality Index Models Using Machine Learning Approaches. *Water Res.* **2023**, *243*, 120337. [CrossRef]
63. Abbasi, T.; Abbasi, S.A. *Water Quality Indices*; Elsevier: Amsterdam, The Netherlands, 2012.

64. Tyagi, S.; Sharma, B.; Singh, P.; Dobhal, R. Water Quality Assessment in Terms of Water Quality Index. *Am. J. Water Resour.* **2020**, *1*, 34–38. [\[CrossRef\]](#)
65. Akter, T.; Jhohura, F.T.; Akter, F.; Chowdhury, T.R.; Mistry, S.K.; Dey, D.; Barua, M.K.; Islam, M.A.; Rahman, M. Water Quality Index for Measuring Drinking Water Quality in Rural Bangladesh: A Cross-Sectional Study. *J. Health Popul. Nutr.* **2016**, *35*, 4. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Zheng, H.; Hou, S.; Liu, J.; Xiong, Y.; Wang, Y. Advanced Machine Learning and Water Quality Index (WQI) Assessment: Evaluating Groundwater Quality at the Yopurga Landfill. *Water* **2024**, *16*, 1666. [\[CrossRef\]](#)
67. Bilgin, A. Evaluation of Surface Water Quality by Using Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI) Method and Discriminant Analysis Method: A Case Study Coruh River Basin. *Environ. Monit. Assess.* **2018**, *190*, 554. [\[CrossRef\]](#)
68. Rahman, A.S.; Rahman, A. Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water* **2020**, *12*, 781. [\[CrossRef\]](#)
69. Nejatian, N.; Yavary Nia, M.; Yousefyani, H.; Shacheri, F.; Yavari Nia, M. The Improvement of Wavelet-Based Multilinear Regression for Suspended Sediment Load Modeling by Considering the Physiographic Characteristics of the Watershed. *Water Sci. Technol.* **2023**, *87*, 1791–1802. [\[CrossRef\]](#)
70. Li, X.; Ding, J.; Ilyas, N. Machine Learning Method for Quick Identification of Water Quality Index (WQI) Based on Sentinel-2 MSI Data: Ebinur Lake Case Study. *Water Supply* **2021**, *21*, 1291–1312. [\[CrossRef\]](#)
71. Sakaa, B.; Elbeltagi, A.; Boudibi, S.; Chaffai, H.; Islam, A.R.M.T.; Kulimushi, L.C.; Choudhari, P.; Hani, A.; Brouziyne, Y.; Wong, Y.J. Water Quality Index Modeling Using Random Forest and Improved SMO Algorithm for Support Vector Machine in Saf-Saf River Basin. *Environ. Sci. Pollut. Res.* **2022**, *29*, 48491–48508. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Lv, M.; Niu, X.; Zhang, D.; Ding, H.; Lin, Z.; Zhou, S.; Zhu, Y. A Data-Driven Framework for Spatiotemporal Analysis and Prediction of River Water Quality: A Case Study in Pearl River, China. *Water* **2023**, *15*, 257. [\[CrossRef\]](#)
73. Fertikh, S.; Boutaghane, H.; Boumaaza, M.; Belaadi, A.; Bouslah, S. Assessment and Prediction of Water Quality Indices by Machine Learning-Genetic Algorithm and Response Surface Methodology. *Model. Earth Syst. Environ.* **2024**, *10*, 5573–5604. [\[CrossRef\]](#)
74. Avand, M.; Janizadeh, S.; Tien Bui, D.; Pham, V.H.; Ngo, P.T.T.; Nhu, V.-H. A Tree-Based Intelligence Ensemble Approach for Spatial Prediction of Potential Groundwater. *Int. J. Digit. Earth* **2020**, *13*, 1408–1429. [\[CrossRef\]](#)
75. Namous, M.; Hssaisoune, M.; Pradhan, B.; Lee, C.-W.; Alamri, A.; Elaloui, A.; Edahbi, M.; Krimissa, S.; Eloudi, H.; Ouayah, M.; et al. Spatial Prediction of Groundwater Potentiality in Large Semi-Arid and Karstic Mountainous Region Using Machine Learning Models. *Water* **2021**, *13*, 2273. [\[CrossRef\]](#)
76. Sheikh Khozani, Z.; Iranmehr, M.; Wan Mohtar, W.H.M. Improving Water Quality Index Prediction for Water Resources Management Plans in Malaysia: Application of Machine Learning Techniques. *Geocarto Int.* **2022**, *37*, 10058–10075. [\[CrossRef\]](#)
77. Saroja; Haseena; Dharshini, S. Deep Learning Approach for Prediction and Classification of Potable Water. *Anal. Sci.* **2023**, *39*, 1179–1189. [\[CrossRef\]](#)
78. Gao, Z.; Chen, J.; Wang, G.; Ren, S.; Fang, L.; Yinglan, A.; Wang, Q. A Novel Multivariate Time Series Prediction of Crucial Water Quality Parameters with Long Short-Term Memory (LSTM) Networks. *J. Contam. Hydrol.* **2023**, *259*, 104262. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.