Cluster Computing (2025) 28:598
https://doi.org/10.1007/s10586-025-05221-3

Advancements in water quality prediction: a practical review of machine learning and deep learning approaches

Marwah A. Helaly[1] · Sherine Rady[1] · Mohamed Mabrouk[1] · Mostafa M. Aref[1] · Sebastian Villarroya[2] · Jose M. Cotos[2] · David Mera[2]

Received: 28 November 2024 / Revised: 3 February 2025 / Accepted: 19 February 2025 / Published online: 30 August 2025
© The Author(s) 2025

Abstract
Water quality plays a pivotal role in ensuring the safety and sustainability of water resources, with significant implications for environmental protection, public health, and various industrial applications. This paper presents both a review of related state-of-the-art works and an implementation and application of adapted versions of these related works for predicting water quality parameters on a new water dataset from Galicia, Spain. The reviewed studies encompass a range of predictive models applied to diverse water quality parameters, including dissolved oxygen levels, pH levels, and other complex water parameters. These models include various machine learning and deep learning methods such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTMs. This research contributes by implementing various models on the dataset and experimentally demonstrating the impact of key factors on model performance. These factors include model sophistication, imputation techniques, recurrent architectures, and customized approaches for water quality prediction using deep learning. Notably, K-Nearest Neighbors (KNN) imputation enhances performance by preserving local data relationships, while noise filtering further improves predictive accuracy. Additionally, we observe that smaller batch sizes and learning rates lead to better generalization in sparse datasets, outperforming traditional approaches. The conclusions are guided by comparing the performance of all models on the Galician dataset using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). This paper provides the first DL-based water quality analysis for Galicia, emphasizing the need for regional model adaptation. Our results guide future research directions, including the exploration of Transformer-based architectures for time-series data, more sophisticated feature selection techniques, and neural-network-based imputation strategies to enhance data completeness.

Keywords Water quality prediction · Deep learning · Machine learning · Data imputation · Convolutional neural network · Recurrent neural network

1 Introduction

Water is essential for all living beings on Earth. In 2015, the United Nations released a 2030 Agenda for Sustainable Development, announcing an urgent call for all countries of the world to collaborate to provide peace and prosperity for the human race and the planet. Seventeen interlinked Sustainable Development Goals (SDGs) were proposed. Among these SDGs, Goal 6—clean water and sanitation—directly impacts Goal 3 (good health and well-being) and Goal 9 (industry, innovation, and infrastructure) [21]. The United Nations listed water quality improvement as a top priority [7], as water quality prediction is crucial for identifying risks related to pollution and environmental degradation, enabling informed decision-making and resource management [15]. Clean water resources are critically required for drinking, agriculture, and industrial sectors.

✉ Marwah A. Helaly
marwah.ahmad.helaly@cis.asu.edu.eg

[1] Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

[2] COGRADE Research Group, Department of Electronics and Computing, University of Santiago de Compostela, Santiago, Spain

Unfortunately, many of the Earth's resources are heavily polluted [4]. Ensuring high water quality reduces the need for water treatment and mitigates the negative effects of poor-quality water on living beings and agriculture [21]. The sector in which the water is intended to be used sets the different water indicators that should be considered when assessing water quality. A common method to assess water health is calculating the Water Quality Index, WQI [20]. The WQI is a simple dimensionless number that combines multiple physical, biological, and chemical water quality parameters together. This provides a simple method to monitor water quality and reduces data storage needs [5]. Most national water quality standards concentrate heavily on *biological* indicators such as Total Coliform (TC), Fecal Coliform (FC), Fecal Streptococcus (FS), and the FC/FS ratio—but research has proved that other indicators are also just as crucial. These other indicators include *physical* indicators such as water temperature, total suspended solids (TSS), total dissolved solids (TDS), and turbidity; *chemical* indicators such as pH, alkalinity, biochemical (biological) oxygen demand (BOD), chemical oxygen demand (COD), NO_3, PO_4, Fe, Mn, Cd, and Zn; and *time* indicators [4, 6, 22].

This paper is organized as follows: first, we briefly outline the objectives of this work, followed by a review of recent studies on the same research topic. In Sect. 4, we describe the various model structures and configurations implemented in this study, detailing any adaptations made and the performance metrics used. Section 5 provides an in-depth look at our dataset, including necessary imputation techniques, preprocessing methods, and an analysis of relationships between data points. Finally, the results are presented and discussed, wrapping up with the paper conclusions.

2 Problem formulation

This paper provides both a review of recent studies on water quality prediction across different datasets and an implementation of these studies on a single dataset from Galicia, Spain. This approach enables a practical comparison of various state-of-the-art models on a new dataset. The Machine Learning (ML) and Deep Learning (DL) models from these works are implemented, with slight adaptations to suit our dataset. This study includes traditional ML techniques, conventional DL approaches, and cutting-edge DL methods. Additionally, two data imputation techniques are applied and compared to evaluate their impact on the predictive models. Consequently, various combinations of models and imputation techniques are explored to identify key factors that experimentally enhance water quality prediction performance. Their evaluation is provided with various performance metrics and also training/validation loss graphs.

3 Literature review

Studies have shown that there are many ways to predict water quality; some predict water quality index, some predict water quality classes, and others predict individual water quality parameters. We will group recent research based on the type of predictions they make.

3.1 Predicting water quality *Indices*

The most recent of such research in Egypt was [16] on the Bahr El-Baqr water resource for irrigation. The authors mentioned three possible harms that result from poor quality irrigation water: salinity hazard, sodicity hazard, and toxicity hazard. Experimenting on 105 water samples for training and testing collected during July 2020 from Bahr El-Baqr, chemical components existing in the water such as sodium *Na*, magnesium *Mg*, and potassium *K*, were used to calculate different water quality criteria. These criteria included the Kelly ratio (KR), the sodium soluble percentage (SSP), the potential of salinity (PS), the permeability index (PI), the sodium adsorption ratio (SAR), and the residual sodium carbonate (RSC). To predict the irrigation water quality index (IWQI), they leveraged both ML models and Regression models. Stepwise regression and Support vector machine produced the best results. Following is [17] who predicted the WQI guided by a hyperparameter grid-search using a support vector regressor (SVR), KNN, Multi-layer perceptron (MLP) regressor, and Decision Tree (DT) models. Also using mean imputation, the MLP regressor providing the highest R^2 with 99.8%, showing the effectiveness of using neural networks.

3.2 Predicting water quality *classes*

In [20] data was explored from both 5 (small experiment) and 19 (large experiment) stations from the Klang River Basin—one of the most polluted rivers in Malaysia. Six water quality indicators were selected in this study to calculate the WQI to use for classification into one of five different water quality classes: dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), pH, suspended solids (SS), and ammoniacal nitrogen (NH_3). The five water quality classes ranged from *Class I—no treatment necessary and very sensitive aquatic species*, to *Class IV—irrigation water*. Three types of classification models were studied, where a deep learning (DL) model performed best with 87.8% accuracy

on the large experiment, a random forest (RF) model performed best with 89% accuracy on the small experiment, and a decision tree (DT) classification models performed the least. However, the DL model exhibited strong stability against non-linear data and varying amounts of data, followed by the RF model. A CNN was also experimented for water class prediction from a *Clean Water* class to an *Unclean Water* class by [3], where the CNN was superior in performance to a KNN model, SVM, and Naive-Bayes model. Also, [17] predicted the Water Quality Class (WQC) into one of three classes guided by a hyperparameter grid-search using Random Forest (RF), Gradient Boosting, Extreme Gradient Boosting, and AdaBoost models. Using mean imputation, the Gradient Boosting model giving the best accuracy of 99.5%.

3.3 Predicting individual water quality *parameters*

Aligning with the SDGs, another Egyptian study attempted the future prediction of individual water quality parameters of the Wadi El-Ryan Upper Lake in 2030 using 1056 samples from seven locations. Three adaptive neuro-fuzzy inference models considered four different water parameters: BOD, COD, ammonia NH_4, and nitrate NO_3 [1]. It was concluded that their second model, which had the largest number of nodes (1450 nodes), performed best with the smallest average MAE of 0.474. However, one of the main issues of this work is that it assumed stable circumstances for the lake, and did not take into consideration any severe changes that may happen that can affect the lake, such as human behavior or weather spikes.

A study of the water quality of the Yellow River - a significant river in China—was conducted by [23] using a multi-task deep learning (DL) model on 36 months of monitored water data. In multitasking, multiple related tasks are learned concurrently in parallel, sharing underlying features and knowledge. The authors divided the section of the river they studied into four sub-sections, where all sections were learned in parallel. Six water quality parameters were considered: water pH, ammonia nitrogen concentration *NHN*, potassium permanganate index $KMnO_4$, chemical oxygen demand (COD), total nitrogen TN, and total phosphorus TP. The multitasking was performed on a hybrid Convolutional Neural Network (CNN) and a Long-Short-Term-Memory (LSTM) model. The CNN was used to extract salient features from the data, while the LSTM was used to learn the long-term dependencies in the water samples. This combination proved much more efficient in comparison to using solely a CNN or an LSTM, decreasing the mean square error (MSE) and root mean square error (RMSE) of the model in comparison to similar models by 13.2% and 15.5%, respectively. Aside from performance, the proposed model also provided better stability and generalization.

Another study was performed on Kastoria Lake in Greece by [10] to control water quality for irrigation. The main goal of the study was to predict the value of dissolved oxygen (DO) in the water, where the water samples from four distributed stations included data regarding the dissolved oxygen, as well as (1) chlorophyll-a, (2) pH, (3) temperature, (4) conductivity, (5) turbidity, (6) ammonia nitrogen NH_4, and (7) nitrate nitrogen NO_3. Other parameters were also included because the literature mentioned that they are also important factors. Standard feed-forward deep neural networks (DNN) with different structures were used. The best network provided a Nash–Sutcliffe model efficiency coefficient (NSE) greater than 0.89.

India is a developing country, and therefore the quality of water sources used for agriculture is especially critical. The water quality factors of five different locations in the Yamuna River, India, over a 6-year period (2013 to 2019) are forecasted by [11] using a SVR, RF, artificial neural network, LSTM, and CNN–LSTM models—but are all outperformed by a Bi-directional LSTM model with the smallest COD MSE of 0.015, and smallest BOD MSE of 0.107. Water parameters considered are such as water temperature, dissolved oxygen, pH, free ammonia *AMM*, COD, BOD, total faecal coliform *TC*, faecal coliform *FC*, and conductivity. A conventional CNN also performed quite well for the prediction of the pH level on water from Atlanta, Georgia in the United States of America [2]. Using variations of four water quality parameters as input including temperature, specific conductance, and the volume of dissolved oxygen, the two-dimensional convolutional network provided low prediction errors.

Reference [4], implemented a hybrid model that merges an LSTM encoder–decoder neural network and a Savitzky–Golay filter in water data from Beijing, China. The data comprised of over 10,000 samples collected every 4 h over a time span from April 2014 to October 2018. With DO and COD as the main water quality parameters, the Savitzky–Golay filter was used to eliminate any noise in the time series data whilst experimenting with several window sizes ranging from 3 to 15. In comparison to a standard artificial neural network and LSTM which produced errors in the ranges of -2 to 2, the proposed filter-LSTM model produced very small error values in the range between -1 and 1.

Finally, the quality of raw drinking water in Oslo and Bergen in Norway was observed by [22]. Water quality indicators that were considered in this study included physical indicators such as water color and turbidity, chemical indicators such as pH and alkalinity, several biological bacteria indicators, and a time indicator. A Pearson's correlation analysis between the different

quality indicators were made to explore possible existing relationships between them, and therefore be able to predict a quality indicator from another. Two prediction models—an adaptive learning rate backpropagation and a random forest model. The random forest model provided lower RMSE and mean absolute error (MAE) than the BP network and other related models. However, the BP network provided higher stability as the data size grows larger but takes a larger training time.

4 Methods

4.1 Models

In this section, we describe the models from the related works that were selected and implemented in our experiments. Due to limitations within our dataset, certain related works were either slightly modified or excluded from our experiments. For instance, some authors included the COD parameter in their input features, so this input was excluded from our model implementations. In cases where related works' models attempted to predict the COD, we had to omit those models from our study altogether.

We adopted all the hyperparameters used by the original authors in the related works considered here, applying them to our dataset-with various imputation methods. Later, we will discuss slight adaptations made to the models during experimentation. An early-stopping mechanism with a patience value of 5 was used to prevent unnecessary iterations. The dataset was split into 80% for training, 10% for validation, and 10% for testing, with shuffling applied to ensure randomness in the dataset distribution. No cross-validation was performed. Additionally, despite some original papers not including the Date feature, we integrated it into every model in our study for consistency across analyses. We implemented these models using Python 3.9 with the Scikit 1.0.2 and Keras 2.15 frameworks.

Next, we delve into the specifics of the models from related works that will be implemented in this study. In a study by [16], a diverse set of seven models was employed to predict six essential irrigation water quality criteria. These models comprised three machine learning (ML) algorithms—namely, Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and Random Forest (RF)—and four regression techniques, including Stepwise Regression (SW) using a backward `scikit-learn SequentialFeatureSelector`, Principal Components Regression (PCR), Partial Least Squares Regression (PLS), and Ordinary Least Squares Regression (OLS). The objective was to predict key criteria such as soluble sodium percentage (SSP), sodium adsorption ratio (SAR), potential of salinity (PS), and Kelly's ratio (KR), among others. We implemented these models on our dataset and also enhanced the predictive power of these models by incorporating additional parameters as exploratory input variables, leveraging the entirety of available data in our dataset. However, despite our comprehensive approach, certain crucial parameters such as Carbonate (CO_3) and Bicarbonate (HCO_3) were notably absent from our dataset. As a result, our models were unable to accurately predict parameters like Residual Sodium Carbonate (RSC) and Permeability Index (PI), highlighting the importance of data completeness in achieving robust predictive outcomes.

Building on the work of [10], who employed traditional neural networks for regression (mathematical foundations are detailed in [12]), we explored networks of varying sizes and configurations. These networks were characterized by two hidden layers, each containing either 32 or 64 nodes. Specifically, they observed four architectures, named as follows:

- 4–32–32–1: 4 inputs, two consecutive 32-node hidden layers, 1 output node,
- 4–64–64–1: 4 inputs, two consecutive 64-node hidden layers, 1 output node,
- 7–32–32–1: 7 inputs, two consecutive 32-node hidden layers, 1 output node,
- 7–64–64–1: 7 inputs, two consecutive 64-node hidden layers, 1 output node.

Remaining faithful to the methodology established by the original authors, in our implementations we adhered to specific training parameters. This included conducting training over 1000 epochs, utilizing an RMSProp optimizer with a learning rate of 0.001, and implementing Rectified Linear Unit (ReLU) activation functions. Additionally, we adopted a batch size of 1 to ensure optimal model convergence during training. Furthermore, the approach necessitated the selection of input parameters crucial for predicting dissolved oxygen (DO), a vital water quality indicator. Two sets of input parameters were considered: one comprising pH, conductivity, water temperature, and nitrate NO_3, while the other included additional variables such as chlorophyll-a, turbidity, and ammonia NH_4 alongside the aforementioned parameters. This meticulous selection aimed to capture a comprehensive range of factors influencing dissolved oxygen levels, thereby facilitating more accurate predictions.

The study by [4], which focused on deep learning methodologies, motivated our investigation into a Long Short-Term Memory (LSTM) based approach for our water quality prediction. We incorporated both the conventional LSTM architecture into our study, and the LSTM model without the Savitzky–Golay filter. This comparative analysis aimed to discern the impact of noise filtering on the

predictive accuracy of the model, therefore offering valuable insights into the effectiveness of such preprocessing techniques in water quality prediction tasks.

Next was the investigation of the impact of employing a simple Convolutional Neural Network (CNN) architecture on a water quality dataset to predict pH levels [2]—mathematical foundations are details in [19]. Their dataset encompassed key parameters such as water temperature, dissolved oxygen, pH, and conductivity, each represented by minimum, maximum, and mean values. Notably, they employed 2D convolutional layers, where the second dimension was derived from organizing the data according to the stations from which they originated. Inspired by their work, we adopted a similar CNN architecture and hyperparameters in our study. Instead of utilizing 2D convolutional layers, we opted for one-dimensional layers. This modification ensures conformity with the architectural patterns employed in the other studies mentioned in this paper, facilitating accurate comparison of predictive performance across different methodologies.

We now delve into a popular area of research that has garnered significant attention in recent years: attention mechanisms. In this model we combine Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, enriched with an attention mechanism [25]. This innovative approach was originally applied by the authors to datasets sourced from the Beilun Estuary in China, focusing specifically on predicting pH and Ammonium Nitrogen (NH_3N). In our own investigation, we sought to replicate and extend upon their work by testing the efficacy of their proposed model with and without the attention mechanism. However, due to limitations in data availability, our study was constrained to utilizing only a subset of the variables used by the authors, comprising pH, DO, Ammonium Nitrogen (NH_3N), air temperature, and water temperature.

The last experimental model [11] featured two convolutional layers, each comprising 250 nodes, followed by a Bi-LSTM layer instead of a conventional LSTM layer. Our adaptions to this model include selecting an Adam optimizer instead of a Bayesian optimization like the original authors. For input data, nine parameters were considered: water temperature, Dissolved Oxygen (DO), pH, free ammonia, Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), total coliform, faecal coliform, and conductivity. These parameters were used to predict Biochemical Oxygen Demand (BOD). Due to the absence of standard BOD measurements in our dataset, we opted to utilize BOD-5 as a substitute. This decision ensures consistency in the predictive modeling framework, despite the slight variation in parameter specifications. Additionally, adjustments were made to the network architecture, such as removing the fully-connected layer and softmax function, to better suit the predictive task at hand. The mathematical foundations of both LSTMs and BiLSTMs are detailed in [18].

4.2 Performance metrics

We selected the performance metrics widely used in related works to evaluate the machine and deep learning models [9], namely Root-Mean Square (RMSE), MAE (Mean-Absolute Errors), and the Coefficient of Determination R^2. In the following equations, these variables are used: n represents the total number of samples in the dataset, i is the index of a single data sample in the dataset, y is the actual truth value of a data sample, and $\hat{y}_i$ is the predicted value of that data sample.

4.2.1 RMSE

The Root Mean Square Error (RMSE) serves as a metric for quantifying the average magnitude of discrepancies between predicted and actual values within a regression framework. Mathematically, RMSE is computed by extracting the square root of the mean of the squared deviations between predicted and observed values, as shown in Eq. (1). Its sensitivity to outliers stems from the squaring operation, accentuating larger deviations and thereby amplifying the impact of extreme values. Consequently, higher RMSE values correspond to larger prediction errors, while an RMSE of 0 denotes ideal model performance, indicating precise predictions without any discrepancy.

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

4.2.2 MAE

The Mean Absolute Error (MAE) is another commonly used measure to evaluate how well a regression model performs. It calculates the average absolute difference between predicted and actual values, as shown in Eq. (2). In simpler terms, MAE is found by averaging the absolute differences between predicted and actual values. Unlike RMSE, MAE doesn't square the errors, so it's less affected by extreme values. Lower MAE values mean the model is performing better, and a value of 0 means the predictions are perfect.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (2)$$

4.2.3 R^2

The Coefficient of Determination (R^2) is a metric that measures the proportion of variability in a dependent variable using the independent variables in a regression model. It serves as an indicator of how well the model fits the data. If an R^2 value ranges from 0 to 1, then 0 suggests that the model fails to capture any variability in the target variable, while 1 indicates a perfect correspondence between the model's predictions and the observed data. Additionally, R^2 can assume negative values if the model performs worse than a simple horizontal line. Higher R^2 values signify superior model performance, with a value of 1 signifying an optimal fit of the model to the dataset. The formula for calculating R^2 is presented in Eq. (3), where $\bar{y}$ is the mean of the actual truth values in the dataset y_i.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (3)$$

5 Dataset and materials

5.1 Dataset

The dataset comprises approximately 4000 data samples collected from water reservoirs in Galicia, a region of northwest Spain, spanning the time period from April 2010 to August 2023. Each sample encompasses a maximum of 76 features, reflecting various water quality parameters. Notably, the dataset exhibits sparsity, with certain parameters exhibiting higher prevalence than others. Included among the features are fundamental water quality indicators such as time, pH levels, water temperature, conductivity, nitrate concentrations, presence of *Escherichia coli* bacteria, and levels of Dissolved Oxygen (DO). No specific feature selection method or model was used—the feature selection process was simply guided by the features used in the related works whose models are mentioned in this work.

Sampling was conducted across 58 distinct locations within Galicia, encompassing three significant sections of the reservoir: the dam, core, and tail. Furthermore, samples were collected at varying depths, including surface, medium depth, and maximum depth levels.

5.2 Data preprocessing

In the context of data preprocessing, categorical variables, such as color and climatology, were numerically encoded. The data was lowercased, invalid values and characters were removed, accent characters were replaced, related columns were merged, and the data was sorted by date. All values were standardized to fall within the range of [0, 1].

Then, a thorough analysis of the dataset revealed variations in the availability of values for each parameter. Consequently, parameters with limited utilization and lower population densities were eliminated, guided by considerations of their relevance to environmental monitoring and consistent with parameters featured in related research literature. As a result, the dataset was refined to encompass 57 parameters across exactly 4228 samples.

Water quality datasets are usually incomplete, often containing many missing values because not all water quality parameters can be measured at every location and time point. Because this holds true for this study as well, we address this issue by next examining the data imputation methods we applied to our Galician dataset.

5.3 Filling missing values

Missing values in water quality datasets are very common. Due to the fact that not all stations at all times have the availability of capturing the entire set of selected information, this causes gaps in the data. There are three types of missing values: (1) Missing Completely at Random (MCAR), where the probability of a missing value is independent of any other variables; (2) Missing at Random (MAR), where missing values depend on observed data but not the missing data itself; and (3) Missing Not at Random (MNAR), which is when missing values are dependent on the values of other features [13].

Missing values in ML are crucial and can greatly degrade the performance of the models, causing the model to falsely bias or reach incorrect conclusions. Consequently, researchers have studied the effects of many solutions, ranging from completely removing the missing data to more sophisticated data imputation methods—which is the data science study of how to fill in missing values—over the years. With water quality data, completely removing rows with missing values is usually out of question, since water samples can have many features and at least one parameter is usually missing. Therefore, research usually is concerned with imputation methods, have been categorized into three groups: statistical methods, model-based methods, and neural network-based methods [26], which get more sophisticated respectively. One of the main concerns of data imputation is to cause the least amount of data distortion possible.

The authors of [24] studied multiple individual statistical imputation methods for water quality datasets: mean, median, random values, and finally a single k-means model (KMM) to predict the missing values. Dataset size was discovered to influence the effectiveness of an imputation method on a dataset, where the KMM and median methods

provided the best test performance. Their study concentrated largely on datasets where the missing data was mainly in a few specific columns, and not many and random missing values from a number of columns.

In this work we imputated the data with two widely-used and commonly adopted methods in related research: once with the median method due to its simplicity and decreased sensitivity to outliers and once with the K-means method, leaving neural-network based methods for future work. The median imputation method works by taking every feature (i.e. column) in the dataset and calculating the single median value for the feature. This single value is then repeated to fill all the missing values in that feature. The KNN imputation method works by calculating the *K*-nearest-neighbours on a feature in the dataset, and fills the value of every missing value using its nearest neighbours.

At this stage, the dataset has been fully prepared. For each data imputation method, Fig. 1 exhibits a Principal Component Analysis (PCA) visualization of the datasets, where each dot in the figure represents a dimensionality-reduced sample in the dataset. The colors of the dots do not have a significance in our study. This shows that there are no extreme outliers in either dataset which is an important piece of information when we come to analyze the performance of the implemented ML and DL models in this work.

6 Results

The performance of ML and DL models is significantly influenced by dataset characteristics, model complexity, and training strategies. This section analyzes the results presented in Figs. 2, 3, 3, 4, 5, and 6, Table 2 for the ML models, and Table 3 for the DL models, highlighting key trends and insights related to model performance.

For the DL models, there are two sets of results:

- **Original (O)**: these results use the exact settings (hyperparameters) chosen by the original authors of the models.
- **Adapted (A)**: we improved the performance of these models on our dataset by adjusting some of the hyperparameters (determined by trial and error). The specific changes made are explained later in this section.

The results of the application of traditional machine learning [16] on our dataset for the prediction of various water quality parameters are shown in Table 2. In predicting soluble sodium percentage (SSP) the models were found to be unsatisfactory, with MAE values ranging between approximately 0.1 and 0.27, which could be deemed acceptable. However, both RMSE and R^2 metrics exhibited poor performance across all models. Notably, the dataset imputed using KNN performed better than the one imputed using the Median method. Among the models, PCR exhibited the best performance in terms of RMSE and MAE, albeit with a considerably low R^2 score. In the prediction of sodium adsorption ratio (SAR), the Extreme Gradient Boosting model demonstrated superior performance, particularly when applied to the median-imputed dataset, yielding an RMSE of 0.396 and an MAE of 0.132, and an impressive R^2 value of 0.708. Conversely, for Kelly ratio prediction, Stepwise Regression emerged as the optimal model, achieving an RMSE of 0.389, MAE of 0.102, albeit with a notably low R^2 value of 0.004. Notably, the prediction dataset for the potential of salinity (POS) appeared to be the most straightforward, as both Stepwise Regression and Ordinary Least Squares Regression models

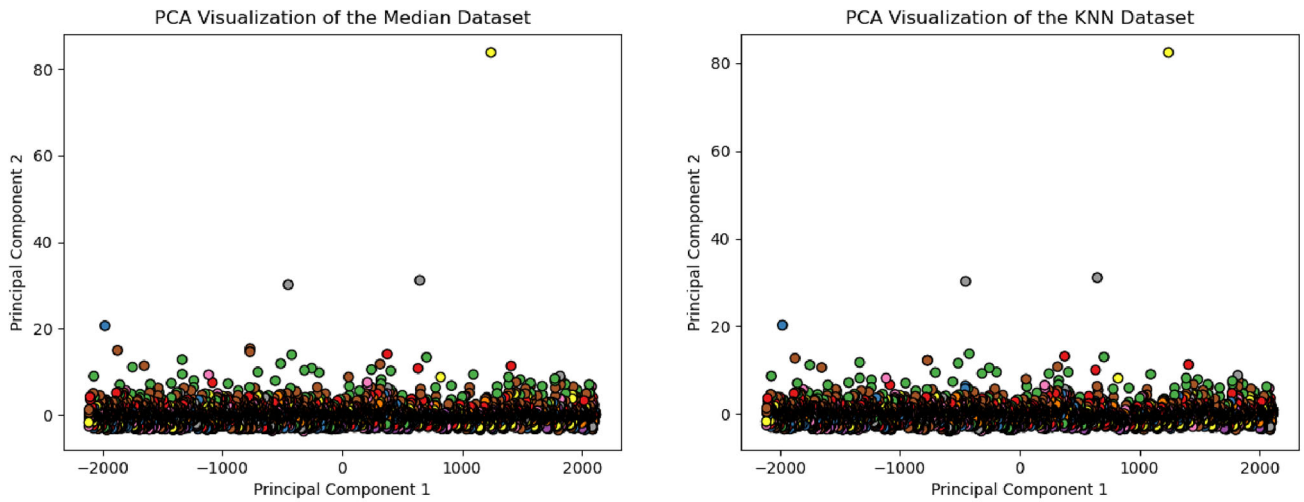
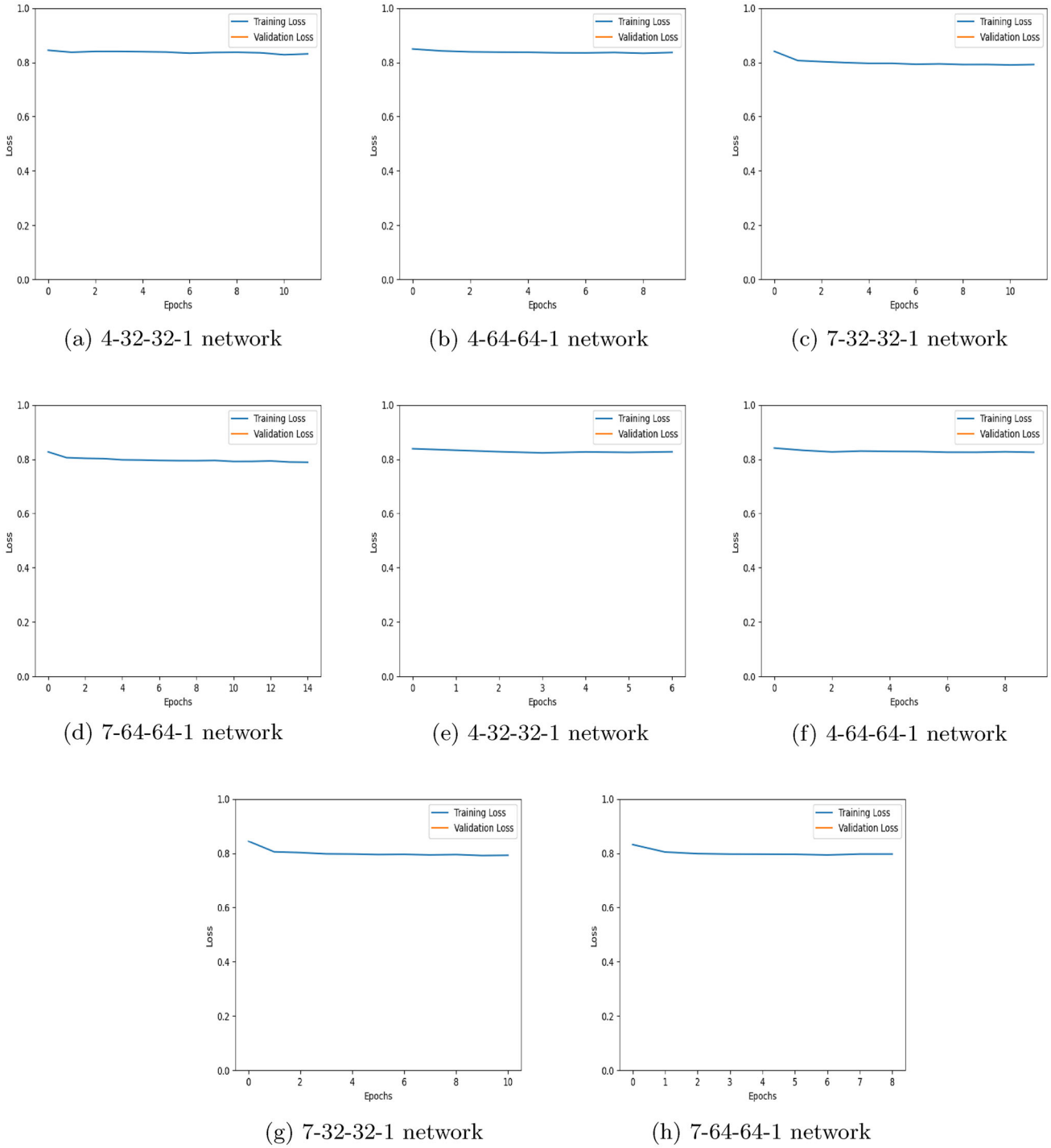


Fig. 1 Principal Component Analysis visualization of the datasets



(a) 4-32-32-1 network (b) 4-64-64-1 network (c) 7-32-32-1 network

(d) 7-64-64-1 network (e) 4-32-32-1 network (f) 4-64-64-1 network

(g) 7-32-32-1 network (h) 7-64-64-1 network

Fig. 2 Validation and training losses using Mean Square Error on our median (subfigures **a**–**d**) and KNN-imputated datasets (subfigures **e**–**h**) for the 4 networks from [10]

achieved perfect performance in this regard. In comparison to the original paper on their own dataset, their models seemed to produce satisfactory R^2 performance, and in some cases the models applied to our datasets would exhibit better RMSE and MAE metrics. It would however be safe to assume that the nature of our dataset is too complex to be efficiently modeled with traditional ML models. The results indicate that these models struggle to generalize well on the sparse and small dataset due to their reliance on handcrafted features and limited ability to capture complex relationships. All traditional models underperform compared to deep learning approaches,

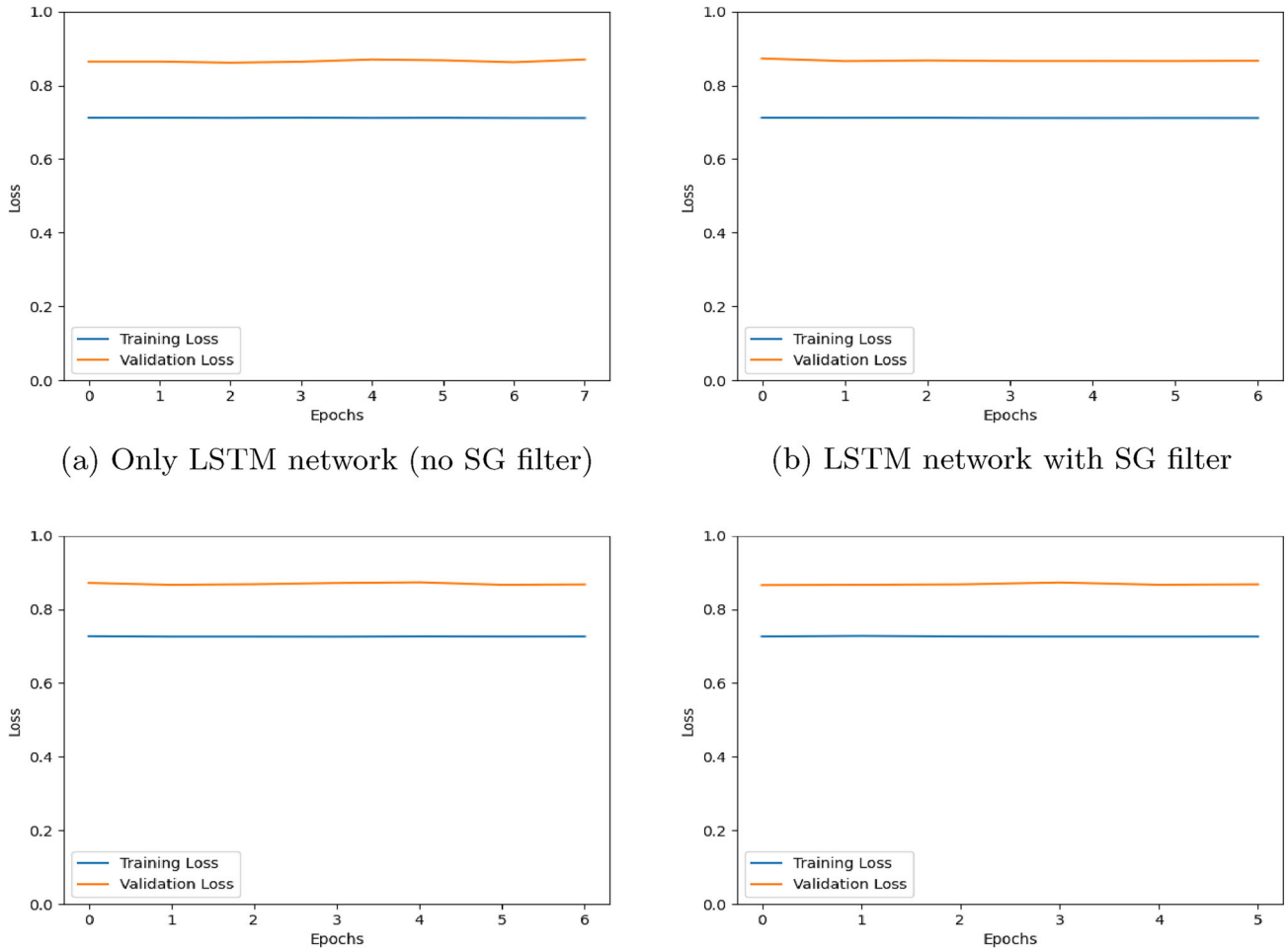
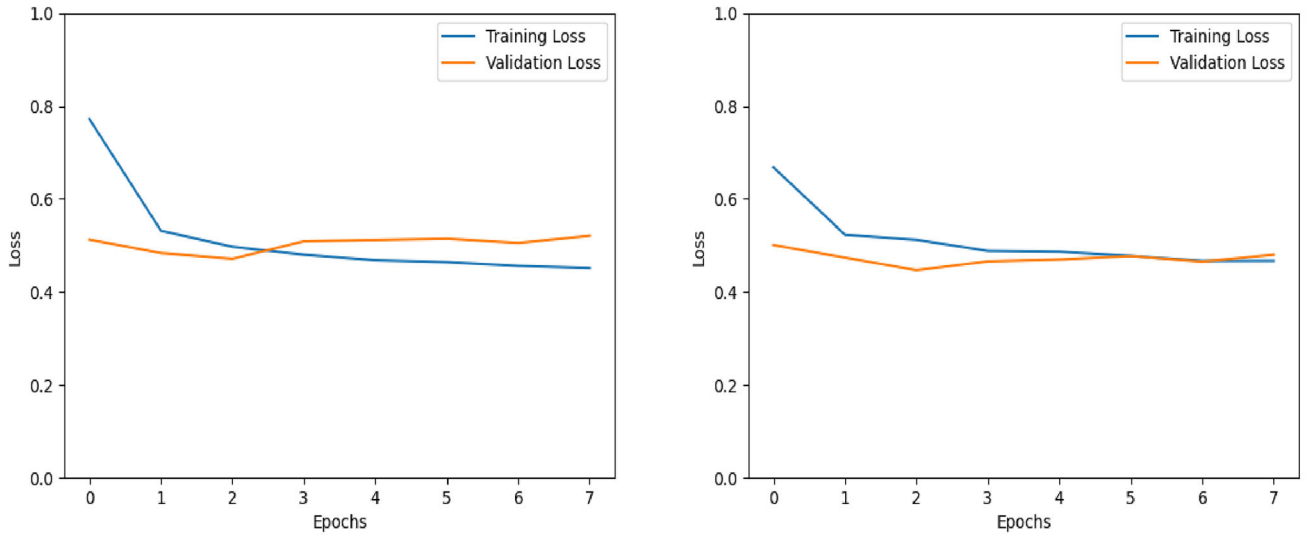


Fig. 3 Validation and training losses using Mean Square Error on our median (subfigures **a**, **b**) and KNN-imputated datasets (subfigures **c**, **d**) for the plain LSTM and SG LSTM networks from [4]

confirming that feature extraction and representation learning play a crucial role in tackling complex datasets like water quality data.

In Fig. 2, the training losses of our dataset applied to the models in [10] all start at slightly more than 0.8 and then stay relatively stable at that range, regardless of the imputation method employed. The validation losses consistently and constantly exceeded 1 for all four models. In other words, the models do not learn. The model may be underfitted and too simple for the task, meaning it cannot capture meaningful patterns in the data. Moreover, the RMSE, MAE, and R^2 metrics as shown in Table 3 exhibited unsatisfactory performance overall, consistently falling within similar ranges. This comparative analysis suggests that a conventional DNN model may lack the complexity required to effectively model our dataset. We attempted to change some essential hyperparameters in the network, such as the learning rate and batch size, but all attempts failed to significantly improve performance. We also experimented with changing the network structure such as 7–8–8–1, 7–32–1, 7–64–1, and 7–8–1, as well as more complex networks such as 7–64–32–8–1, 7–1–6–8–4–1, and 7–64–32–16–8–4–1, while also testing 4 input parameters instead of 7. For all models, the validation errors during training were the same as shown in Fig. 2—all being more than 1, but the testing performance was always more or less the same as those shown in the respective part of Table 3. Given the small dataset size and sparsity, these models do not learn meaningful feature representations, leading to convergence issues and overfitting. Table 3 quantifies this underperformance, showing that DNNs consistently lag behind more specialized architectures.

In the prediction task concerning Dissolved Oxygen (DO), when applied to our dataset the LSTM models presented by [4] showcased performance surprisingly



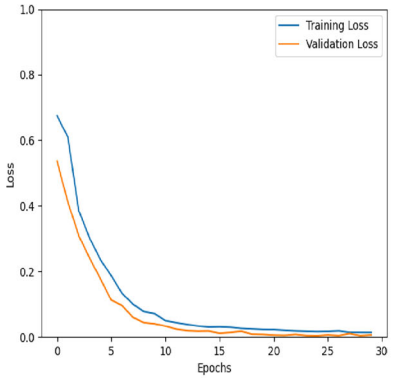
(a) Median-imputated dataset. (b) KNN-imputated dataset.

Fig. 4 Validation and training losses using Mean Square Error on our median and KNN-imputated datasets for the plain CNN from [2]

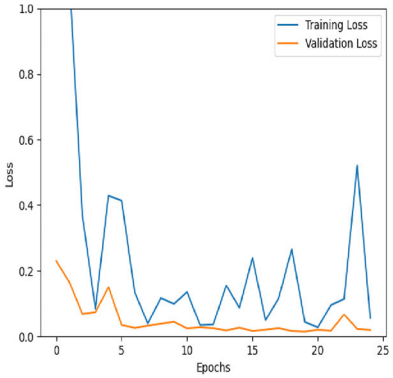
comparable to the traditional DNN models introduced by [10]. In Fig. 3, we observe that the training losses remain stable at around 0.7. The validation loss stays higher at around 0.8, indicating a significant and constant gap between training and validation performance. This also indicates that the underfitted models do not learn and their training and validation losses are constant high losses—similar to [10]. The inclusion of the Savitzky–Golay (SG) filter for noise reduction in the LSTM architecture for the median-imputed dataset resulted in marginally improved performance compared to the traditional LSTM model. This observation suggests that the dataset may not manifest strong temporal characteristics substantial enough to be adequately captured and utilized by the LSTM model. However, changing the batch size to 1 improved the performance of both datasets with and without the Savitzky–Golay (SG) filter. Moreover, we achieved almost perfect performance on the KNN imputed dataset with the Savitzky–Golay (SG) filter.

In the realm of pH level prediction of [2], a straightforward one-dimensional CNN outperformed the LSTM model, exhibiting approximately a 10–15% enhancement in performance. In Fig. 4, the training losses for both figures starts high between 0.7 and 0.8 and decreases steadily over the epochs, indicating that the model is learning. Around epochs 2 and 3, the validation losses increases slightly and fluctuates, which may suggest overfitting beginning to occur. Adapting to our datasets by changing the learning rate to 0.0002 from the original 0.01 and batch size to 8 from 120 gave a slightly better performance, with the median-imputed dataset slightly better than the KNN dataset.

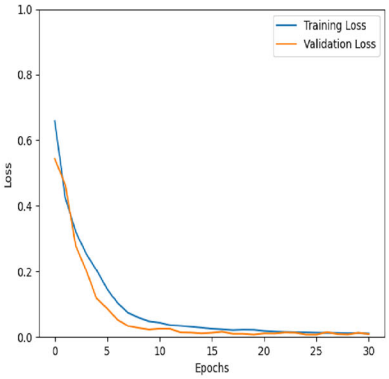
An CNN–LSTM with attention model was proposed by [25] for the tasks of pH and ammonium prediction tasks. We experiment with the model both with and without attention. Figure 5 exhibits all the models successfully learning and converging to different levels. The figures for the Total Ammonium prediction show that the models' learning heavily fluctuate and have more trouble learning than the pH models - regardless of the type of data imputation and whether attention is used in the models or not. For pH prediction, the combined employment of CNN and LSTM models yielded nearly flawless performance, as illustrated in Table 3. Moreover, the incorporation of attention mechanisms further enhanced the predictive capability, particularly with the KNN-imputed dataset exhibiting marginally superior performance compared to the median-imputed dataset. It is noteworthy that there existed discrepancies in the training parameters utilized between this study and the directly previous work research, as outlined in Table 1. Specifically, this study incorporated additional input features such as date, pH level, dissolved oxygen (DO), total ammonium, and air temperature. Any changes we made to the hyperparameters did not improve the performance for pH prediction. Overall, attention was a valuable addition to the model, making the model more stable to differences in hyperparameters. The learning of the pH parameter seemed to be easier than that of Total Ammonium, exhibiting that some parameters are easier to learn than others. Nonetheless, satisfactory performance was still achieved, particularly with attention mechanisms applied to the KNN-imputed dataset, resulting in an RMSE of 0.048, MAE of 0.035, and R^2 value of 0.998. Changing the batch size to 1 instead of 24 on our KNN-imputed



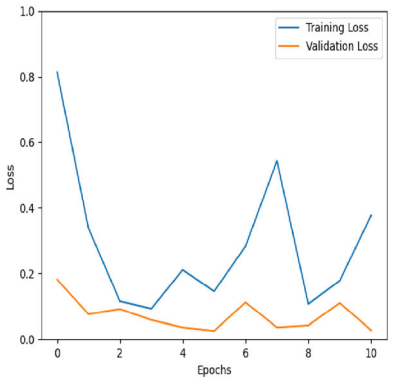
(a) Median-imputated dataset for pH prediction.



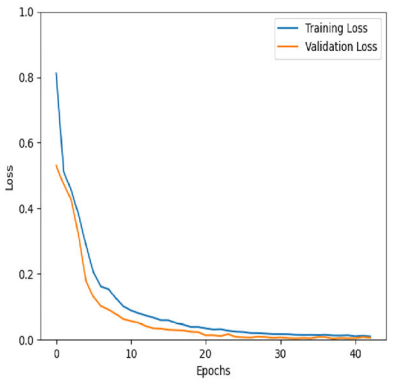
(b) Median-imputated dataset for Total Ammonium prediction.



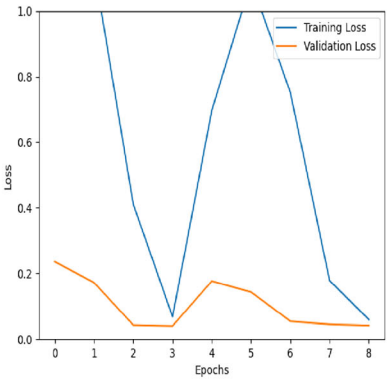
(c) KNN-imputated dataset for pH prediction.



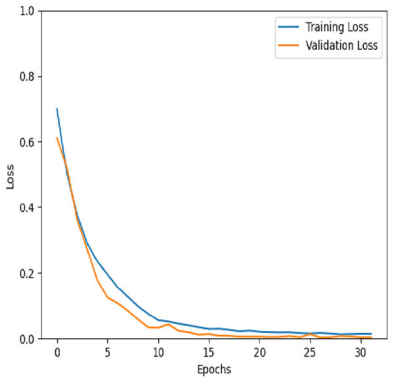
(d) KNN-imputated dataset for Total Ammonium prediction.



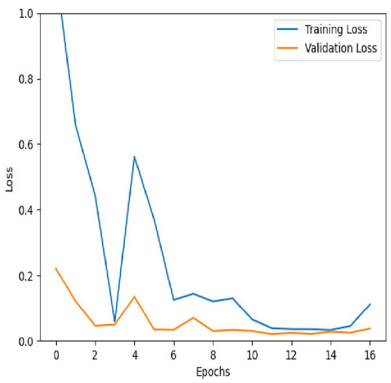
(e) Median-imputated dataset for pH prediction.



(f) Median-imputated dataset for Total Ammonium prediction.

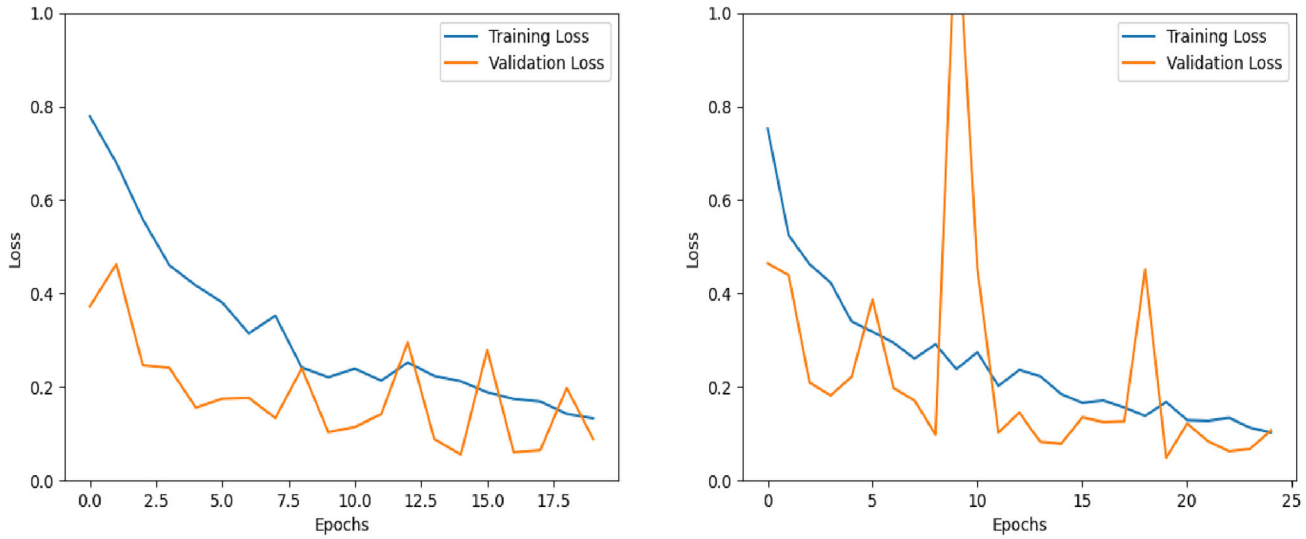


(g) KNN-imputated dataset for pH prediction.



(h) KNN-imputated dataset for Total Ammonium prediction.

Fig. 5 Training and validation losses using Mean Square Error for the Yang CNN–LSTM model without attention (subfigures **a**–**d**) and with attention (subfigures **e**–**h**) on median and KNN-imputated dataset for both pH and Total Ammonium prediction



(a) Median-imputated dataset. (b) KNN-imputated dataset.

Fig. 6 Validation and training losses using Mean Square Error on our median and KNN-imputated datasets for the Khullar CNN–BiLSTM model for pH prediction

Table 1 Original input and output parameters of related works considered in this practical comparative study

Paper	Year	Country	Water source	Input parameter	Output parameter	Model
[16]	2022	Egypt	Bahr El-Baqr	Na, Ca, Mg, K, CO_3, HCO_3, Cl, SO_4	SSP, SAR, RSC, PS, PI, KR	SVM, XGB, RF, SW, PCR, PLS, OLS
[10]	2021	Greece	Kastoria Lake	pH, conductivity, water temperature, NO_3, chlorophyll-a, turbidity, NH_4	DO	Traditional DNNs
[2]	2023	USA	From Atlanta, Georgia	Temperature, dissolved oxygen, pH, conductivity	pH	Traditional CNN
[4]	2021	China	From Gubeikou Town	DO, COD	DO, COD	LSTM + Savgol filter
[11]	2022	India	Yamuna River	Water temperature, DO, pH, free ammonia, COD, BOD, total coliform, faecal coliform, conductivity	BOD, COD	CNN + BiLSTM
[25]	2021	China	Beilun Estuary	pH, DO, COD, NH_3–N, air temperature, atmospheric pressure of weather station, atmospheric pressure of mean sea level, humidity, wind speed, visibility, dew point temperature, rainfall	pH, NH_3N	CNN + LSTM + Attention

dataset gave a near-perfect performance, making the usage of attention fairly invaluable and deteriorated the performance. It can be concluded that the learning and prediction of Ammonium parameter was a more difficult task than that of the pH level parameter.

Finally, for the unique task of predicting DOB-5, the CNN–BiLSTM model proposed by [11] demonstrated satisfactory performance. Figure 6 exhibits the models successfully learning and converging, showing strong performance. Interestingly, the median-imputed dataset exhibited slightly superior performance compared to the KNN-imputed dataset. To improve the performance, we decreased the learning rate from 0.01 to 0.001, used a default Adam optimizer, and drastically decreased the batch size from 120 to 16 gaving us a strong boost in performance for both the median- and KNN-imputed dataset, with the KNN-imputation coming on top with a 0.091 RMSE, 0.057 MSE, and 0.988 R^2 measures.

Table 2 Compound water quality parameter prediction with various ML methods, using both Median- and KNN-imputed data [16]

Model	Paper	Median			KNN		
		RMSE	MAE	R^2	RMSE	MAE	R^2
		Predict Compound—Soluble Sodium Percentage (SSP)					
SVR	[16]	0.768	0.197	0.033	1.473	0.115	0.0
XGB	[16]	0.944	0.235	− 0.539	1.345	0.097	− 0.787
RF	[16]	0.765	0.221	− 0.019	1.579	0.121	− 0.139
SW	[16]	1.028	0.268	0.004	1.481	0.1	− 0.001
PCR	[16]	0.787	0.229	0.006	**0.162**	**0.064**	**− 0.088**
PLS	[16]	0.589	0.223	0.004	0.443	0.106	−0.054
OLS	[16]	0.825	0.253	− 0.022	0.354	0.136	− 0.501
		Predict Compound—(SAR)					
SVR	[16]	1.203	0.236	0.218	1.002	0.225	0.247
XGB	[16]	**0.396**	**0.132**	**0.708**	**0.427**	**0.165**	**0.62**
RF	[16]	0.672	0.234	0.137	0.863	0.266	0.258
SW	[16]	0.731	0.386	0.091	0.808	0.312	0.089
PCR	[16]	0.844	0.377	0.079	1.168	0.479	0.029
PLS	[16]	0.737	0.321	0.072	0.719	0.354	0.245
OLS	[16]	0.709	0.372	0.031	0.766	0.373	0.187
		Predict Compound—Kelly's Ratio (KR)					
SVR	[16]	0.657	0.109	0.052	1.603	0.158	0.015
XGB	[16]	1.487	0.136	0.015	1.541	0.141	0.07
RF	[16]	0.698	0.156	− 0.421	1.611	0.157	− 0.003
SW	[16]	0.62	0.119	− 0.002	**0.389**	**0.102**	**0.004**
PCR	[16]	1.235	0.159	− 0.004	0.346	0.098	− 0.007
PLS	[16]	1.023	0.16	0.018	0.477	0.168	−0.016
OLS	[16]	1.231	0.201	− 0.031	0.387	0.187	− 0.57
		Predict Compound—(POS)					
SVR	[16]	1.051	0.12	0.213	0.22	0.084	0.867
XGB	[16]	1.274	0.073	0.318	0.052	0.02	0.991
RF	[16]	0.546	0.275	0.457	1.324	0.254	0.292
SW	[16]	**0.0**	**0.0**	**1.0**	**0.0**	**0.0**	**1.0**
PCR	[16]	0.819	0.35	0.426	1.36	0.388	0.193
PLS	[16]	0.601	0.178	0.789	0.266	0.183	0.805
OLS	[16]	**0.0**	**0.0**	**1.0**	**0.0**	**0.0**	**1.0**

The highest result for each type of experiment is highlighted in bold

7 Discussion

The results show that DL models outperform traditional ML models, especially for complex water quality parameters. The positive impact of noise filtering suggests that data preprocessing techniques significantly improve model performance by mitigating irrelevant information. The improved performance observed with KNN-imputed data highlights the importance of effective imputation strategies. KNN likely preserves more relevant information compared to simpler methods, leading to better model training.

When it comes to hyperparameters, smaller batch sizes and learning rates performed better, suggesting that larger configurations may lead to overfitting. Additionally, the effectiveness of sophisticated optimizers suggests their role in achieving convergence and optimal parameter updates.

Machine learning and traditional deep learning models can generalize under the right conditions [8]. However, our results indicate that even traditional deep neural networks (DNNs) struggle due to dataset sparsity and small size, leading to poor learning and convergence. These models struggle to capture intricate relationships within the dataset. This aligns with previous findings that traditional architectures may not be well-suited for highly sparse and small datasets, where more expressive or specialized architectures are required. However, the effect of feature selection may be studied to potentially improve performance [14].

Table 3 Individual water quality parameter prediction with various DL methods, using both Median- and KNN-imputed data

Model	Paper	Version	Median			KNN		
			RMSE	MAE	R^2	RMSE	MAE	R^2
			Predict Dissolved Oxygen (DO)					
Traditional NN 4–32–32–1	[10]	O	0.884	0.507	0.076	0.86	0.5	0.119
Traditional NN 4–64–64–1	[10]	O	0.872	0.499	0.102	0.879	0.51	0.079
Traditional NN 7–32–32–1	[10]	O	0.85	0.477	0.147	0.854	0.486	0.13
Traditional NN 7–64–64–1	[10]	O	0.853	0.489	0.141	0.852	0.487	0.136
			Predict Dissolved Oxygen (DO)					
LSTM	[4]	O	0.828	0.611	0.012	0.83	0.627	0.008
LSTM	[4]	A	0.073	0.036	0.992	0.084	0.068	0.99
SG LSTM	[4]	O	0.83	0.616	0.007	0.833	0.629	0.002
SG LSTM	[4]	A	0.025	0.021	0.999	**0.015**	**0.007**	**1.0**
			Predict pH					
CNN	[2]	O	0.718	0.513	0.475	0.677	0.476	0.528
CNN	[2]	A	**0.648**	**0.434**	**0.572**	0.655	0.467	0.559
			Predict pH					
CNN–LSTM	[25]	O	0.076	0.059	0.994	0.079	0.055	0.994
CNN–LSTM	[25]	A	0.155	0.054	0.975	0.073	0.045	0.994
CNN–LSTM with Attention	[25]	O	0.067	0.045	0.995	**0.048**	**0.035**	**0.998**
CNN–LSTM with Attention	[25]	A	0.077	0.057	0.994	0.085	0.06	0.993
			Predict Total Ammonium					
CNN–LSTM	[25]	O	0.163	0.086	0.561	0.194	0.106	0.374
CNN–LSTM	[25]	A	0.092	0.082	0.861	**0.069**	**0.044**	**0.921**
CNN–LSTM with Attention	[25]	O	0.215	0.114	0.236	0.184	0.106	0.438
CNN–LSTM with Attention	[25]	A	0.216	0.12	0.228	0.203	0.115	0.318
			Predict DOB-5					
CNN–BiLSTM	[11]	O	0.279	0.194	0.882	0.311	0.254	0.864
CNN–BiLSTM	[11]	A	0.113	0.058	0.981	**0.091**	**0.057**	**0.988**

The highest result for each type of experiment is highlighted in bold

Attention allows models to focus on crucial data points, potentially leading to improved predictions. However, our results show that with some parameters—like Total Ammonium—using Attention does not improve the performance. This is opposite of what happened for the pH level prediction—which improved its results, which most recent studies usually prove. The interesting benefits of attention mechanisms, when and when not to use it, or how to further preprocess a dataset so that it benefits from Attention—encourage further exploration.

Long Short-Term Memory (LSTM) networks perform very well on water quality parameters. This is due to their ability to handle sequential data and this highlights the importance of capturing temporal dependencies for water quality prediction. While Convolutional Neural Networks (CNNs) can be useful for feature extraction, relying solely on them might be insufficient for capturing the complex dynamics of water quality data. In general, CNN-based architectures improve performance by leveraging spatial patterns in data. These models outperform standard DNNs by effectively extracting local dependencies among water quality parameters, making them more resilient to sparsity. BiLSTMs learn from past and future information, making them potentially advantageous for water quality prediction. However, comparing the performance of BiLSTMs to LSTMs (and even CNN–LSTMs) for dissolved oxygen (DO) prediction—suggests a need for deeper investigation into using both architectures.

There is no single best model for predicting water quality parameters. The observed differences in performance across prediction tasks highlight that some parameters are harder to predict than others and that models need to be tailored to the specific parameter being predicted. For example, in a CNN–LSTM model predicting pH versus Total Ammonium, pH was easier to learn.

Our work faced several limitations, mainly revolving around a small dataset size and incomplete water quality data, which resulted in missing values and a sparse dataset. These missing values required data imputation. The imputed values may not always perfectly represent the actual values and are likely to differ from true values. The configuration of our dataset however, brought out how strong some prediction models are in comparison to others.

Further research into feature selection strategies might help determine the most relevant water quality parameters. Exploring sophisticated imputation methods (e.g., neural-network-based models) may yield more meaningful and precise estimates for missing values. The effect of using noise filters on a dataset was also interesting and could be further explored. More sophisticated DL models—such as Transformers—are proven capable of efficiently learning from time-series data. While this study followed related works for performance metrics, further research may explore alternatives like the Nash–Sutcliffe efficiency coefficient for a more comprehensive evaluation. Larger datasets may significantly enhance DL model performance, as they excel with extensive data. Since our dataset's source is only from Galicia, this means the models may not generalize well to other locations. More diverse datasets with additional parameters from different countries could provide tremendous benefits—resulting in a generalized model that can learn from and benefit multiple regions. This could potentially be achieved by adding input parameters to cover multiple locations, as different regions are sensitive to different water parameters due to factors like geography, climate, land use, and pollution sources. Adding a "City" or "Country" parameter to differentiate locations is essential. The application of Transfer Learning (TL) between models trained on datasets from different locations could be considered and would provide valuable insights.

These findings clarify how various factors in ML and DL models impact water quality prediction. This information can guide the development of more accurate and complex systems, like water resource management systems, and help countries create policies for managing water resources. Efficiently predicting specific water quality parameters will enhance monitoring. Furthermore, the observed benefits of attention mechanisms and CNN–LSTM-based models suggest that real-time monitoring systems could be enhanced by implementing such architectures. These systems could provide early warnings of water quality deterioration, enabling more effective water treatment and pollution control strategies. Our results are specific to Galicia, Spain, due to the dataset's geographic scope, which allows for a deeper understanding of water quality levels for Galician communities.

8 Conclusion

This paper presents a study that investigates the application of various machine and deep learning models for predicting water quality parameters. Implemented on a Galician dataset, our findings highlight key considerations for effective model development in this domain. This study highlights the significant potential of deep learning models in predicting water quality parameters, especially when coupled with effective data preprocessing techniques and tailored model architectures. Although challenges such as dataset sparsity and imputation limitations remain, our findings offer a clear direction for future research, including exploring advanced imputation strategies, optimizing attention mechanisms, and leveraging Transfer Learning for model generalization across regions. By expanding datasets and refining model configurations, we can develop more accurate, real-time monitoring systems that support better water resource management and environmental protection. Ultimately, this work contributes to the ongoing efforts to enhance water quality prediction models and informs the development of more effective strategies for managing water resources.

Acknowledgements We would like to thank the Center of Augas de Galicia from the Ministry of Environment and Infrastructure of the Galician Government for providing us with the dataset. This work was collaboratively performed during an Erasmus Scholarship for a PhD Research Visit to the Electronics and Computer Science Department at the University of Santiago de Compostela, Spain.

Author contributions Marwah A. Helaly: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft and review and editing. Sherine Rady: Supervision, writing—review and editing. Mohamed Mabrouk: Supervision, writing—review and editing. Mostafa M. Aref: Supervision, writing—review and editing. Sebastian Villaroya: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, supervision, validation, writing—review and editing. Jose M. Cotos: Conceptualization, data curation, formal analysis, investigation, supervision, writing—review and editing. David Mera: Data curation, methodology, validation, writing—review and editing. All authors read and approved the final manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability The data that support the findings of this study are available upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval and Informed consent Our research does not involve humans, animals, or any biological material. Therefore, ethical approval and informed consent are not required.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abd El-Mageed, A.M., Enany, T.A., Goher, M.E., Hassouna, M.E.: Forecasting water quality parameters in Wadi El Rayan Upper Lake, Fayoum, Egypt using adaptive neuro-fuzzy inference system. Egypt. J. Aquat. Res. **48**(1), 13–19 (2022)
2. Al-Shourbaji, I., Duraibi, S.: IWQP4Net: an efficient convolution neural network for irrigation water quality prediction. Water **15**(9), 1657 (2023)
3. Anand, M.V., Sohitha, C., Saraswathi, G.N., Lavanya, G.: Water quality prediction using CNN. J. Phys. Conf. Ser. **2484**, 012051 (2023)
4. Bi, J., Lin, Y., Dong, Q., Yuan, H., Zhou, M.: Large-scale water quality prediction with integrated deep neural network. Inf. Sci. **571**, 191–205 (2021)
5. Ewaid, S.H., Abed, S.A., Al-Ansari, N., Salih, R.M.: Development and evaluation of a water quality index for the Iraqi rivers. Hydrology **7**(3), 67 (2020)
6. Galal, M., Soliman, A., Kamel, G., Zaher, K., El-Fakharany, Z.: Prediction and assessment of surface water quality effect on groundwater in El-Galuybia, Egypt. J. Eng. Appl. Sci. **67**, 2129–2148 (2020)
7. Giri, S.: Water quality prospective in twenty first century: status of water quality in major river basins, contemporary strategies and impediments: a review. Environ. Pollut. **271**, 116332 (2021)
8. Helaly, M.A., Rady, S., Aref, M.M.: BERT contextual embeddings for taxonomic classification of bacterial DNA sequences. Expert Syst. Appl. **208**, 117972 (2022)
9. Hodson, T.O.: Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. Discuss. **2022**, 1–10 (2022)
10. Karamoutsou, L., Psilovikos, A.: Deep learning in water resources management: the case study of Kastoria Lake in Greece. Water **13**(23), 3364 (2021)
11. Khullar, S., Singh, N.: Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. Environ. Sci. Pollut. Res. **29**(9), 12875–12889 (2022)
12. Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: Exploring strategies for training deep neural networks. J. Mach. Learn. Res. **10**(1), 1–40 (2009)
13. Lee, K.J., Carlin, J.B., Simpson, J.A., Moreno-Betancur, M.: Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. Int. J. Epidemiol. **52**(4), 1268–1275 (2023)
14. Mera, D., Bolon-Canedo, V., Alonso-Betanzos, A.: On the use of feature selection to improve the detection of sea oil spills in SAR images. Comput. Geosci. **100**, 166–178 (2017)
15. Mera, D., Cotos, J.M., Varela-Pet, J., Rodríguez, P.G., Caro, A.: Automatic decision support system based on SAR data for oil spill detection. Comput. Geosci. **72**, 184–191 (2014)
16. Mokhtar, A., Elbeltagi, A., Gyasi-Agyei, Y., Al-Ansari, N., Abdel-Fattah, M.K.: Prediction of irrigation water quality indices based on machine learning and regression models. Appl. Water Sci. **12**(4), 76 (2022)
17. Shams, M.Y., Elshewey, A.M., El-Kenawy, E.S.M., Ibrahim, A., Talaat, F.M., Tarek, Z.: Water quality prediction using machine learning models based on grid search method. Multimed. Tools Appl. **83**(12), 35307–35334 (2024)
18. Siami-Namini, S., Tavakoli, N., Namin, A.S.: The performance of LSTM and BiLSTM in forecasting time series. In: 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 3285–3292. IEEE (2019)
19. Sze, V., Chen, Y.H., Yang, T.J., Emer, J.S.: Efficient processing of deep neural networks: a tutorial and survey. Proc. IEEE **105**(12), 2295–2329 (2017)
20. Tiyasha, T., Tung, T.M., Yaseen, Z.M.: Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. Nat. Resour. Res. **30**(6), 4235–4254 (2021)
21. United Nations: United Nations Sustainable Development Goals. United Nations (2023). https://sdgs.un.org/goals. Accessed 1 Nov 2023
22. Wu, D., Wang, H., Seidu, R.: Smart data driven quality prediction for urban water source management. Future Gener. Comput. Syst. **107**, 418–432 (2020)
23. Wu, X., Zhang, Q., Wen, F., Qi, Y.: A water quality prediction model based on multi-task deep learning: a case study of the Yellow River, China. Water **14**(21), 3408 (2022)
24. Yang, R.: Analyses of approaches to deal with missing data in water quality data set. In: 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022), 2022, pp. 1102–1108. Atlantis Press (2022)
25. Yang, Y., Xiong, Q., Wu, C., Zou, Q., Yu, Y., Yi, H., Gao, M.: A study on water quality prediction by a hybrid CNN–LSTM model with attention mechanism. Environ. Sci. Pollut. Res. **28**(39), 55129–55139 (2021)
26. Zhang, Y., Thorburn, P.J.: Handling missing data in near real-time environmental monitoring: a system and a review of selected methods. Future Gener. Comput. Syst. **128**, 63–72 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

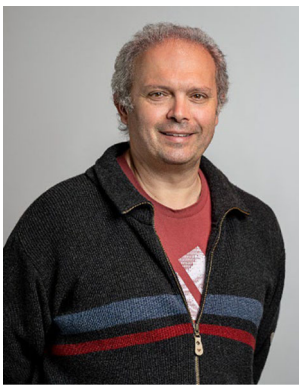
Marwah A. Helaly is a researcher in Artificial Intelligence and Deep Learning with extensive academic research experience. She is currently pursuing a Ph.D. in Computer Science at the Faculty of Computer and Information Sciences, Ain Shams University, where she also earned her BSc and MSc. She is also an Assistant Lecturer at Ain Shams University. Her research focuses on Deep Learning applications in water quality prediction and DNA sequence classification, aiming to drive AI innovation in computational science. She has published research in peer-reviewed journals and conferences, contributing to advancements in AI-driven environmental and biological applications.

Sherine Rady is a Professor at the Faculty of Computer and Information Sciences of Ain Shams University in Cairo, Egypt. She holds a B.Sc. in Electrical Engineering (Computer and Systems), Ain Shams University. She got her M.Sc. in Computer and Information Sciences from Ain Shams University and her Ph.D. from University of Mannheim in Germany. – Prof. Sherine Rady is a DAAD and JICA Alumni and her research interests are Artificial Intelligence, Data Science and Big Data.

Mohamed Mabrouk received the B.Sc. degree in Computer Science from the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. From 2002 to 2006, he was pursuing his M.Sc. at the same faculty. From 2010 to 2013 he was conducting his Ph.D. at the University of Leipzig, Germany. The area of his Ph.D. is Semantic Web and Linked Data. The main focus of his Ph.D. was the efficient extraction of data from Wikipedia and converting this data into (Resource Description Framework) RDF format in order to enable efficient querying of this data via SPARQL. From 2014 to 2016 He was conduction PostDoc research at the University of Amsterdam, the Netherlands. The focus of his research was to model Cloud Computing infrastructure using ontologies and RDF, in order to enable the efficient management and provisioning of the various cloud components, e.g. Virtual Machines, Data storage. Currently, he is an Associate Professor at the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. His research interests include AI, Machine Learning, Semantic Web, Knowledge Graphs, Ontologies, and LLMs.

Prof. Mostafa Aref is a professor of Computer Science, Ain Shams University, Cairo, Egypt. Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT. Research area: Natural Language Processing, Knowledge Representation and Ontology.

Dr. Sebastian Villarroya is an assistant professor at Universidade de Santiago de Compostela. He worked as a research associate at Jacobs University Bremen and Universidade de Santiago de Compostela. He obtained his PhD at Universidade de Santiago de Compostela, researching integrated modeling and analysis of big raster and vector data. Beyond distributed big data analysis, sensor data acquisition systems and big spatial data analytics, he is currently focused on two emerging research fields: integration of machine learning technologies and raster database management systems, and quantum databases.



Jose M. Cotos has been a Professor with the Department of Electronics and Computing, Universidad de Santiago de Compostela, since 1993. He is currently the Coordinator of the Computer Graphics and Data Engineering Research Group. He participated in more than 20 research projects and in more than 50 contracts with companies and institutions, mostly related to the transfer of technology to the business sector. From 2009 to 2013, he was attached to the presidency of a university network for technology transfer, RedEmprendia. In addition, he was the Founding Partner and Administrator of the spin-off Paralaxe, Multimedia and Virtual Systems SL, a spin-off company of the Institute of Technological Research, University of Santiago de Compostela that was dedicated to the development of multimedia and virtual reality computer systems. Currently, he is involved in the implementing of Machine Learning algorithms, applied to industrial processes.