2. REVIEW OF LITERATURE

In recent years, there has been growing concern about the quality of river water, and various methods have been developed to predict it. The prediction of river water quality is an important task in environmental monitoring and management. Machine learning and deep learning approaches have become popular for this purpose due to their ability to handle large amounts of data and complex relationships between variables. In recent years, several studies have focused on using these techniques to predict river water quality index. The literature review shows that machine learning and deep learning approaches have achieved promising results in predicting these indices, outperforming traditional statistical methods in terms of accuracy and speed. However, the choice of features, model selection, and hyperparameter tuning are critical factors that can affect the performance of these models. This chapter provides an overview of the latest research on machine learning and deep learning approaches for river water quality index prediction and identifies the challenges and opportunities for further research.

2.1 STATISTICAL APPROACHES IN WQI PREDICTION

Traditional time series analysis methods were widely used for water quality index prediction. These methods involve analysing historical data to identify patterns, trends, and seasonality in the data. The most common time series approaches used for water quality index prediction include ARIMA, ES, and seasonal decomposition of time series (STL). ARIMA models were widely used for their ability to capture the linear dependencies between time series data, while ES models were useful for capturing non-linear patterns and trends. STL was a robust method that decomposes a time series into seasonal, trend, and residual components, making it useful for analysing complex patterns in water quality index data. The traditional time series approaches were previously applied to predict the water quality index, but their accuracy was found to be unsatisfactory, and they were limited in their ability to handle large volumes of data. Therefore, it is important to carefully select the appropriate time series model and validate its performance on a range of data sets. Some of the statistical approaches that have been used in the past are reviewed and stated below.

Shrestha [49] presented a comprehensive study on the pollution status and water quality of the Ratuwa River and its tributaries in Damak, Nepal. The objective of the research was to analyse and interpret the hydro chemical parameters of the river water and evaluate the percentage contribution to the electrical conductivity (EC) using statistical techniques. The

approach used in the study involved the collection of water samples from different sites along the Ratuwa River and its tributaries. The samples were collected at two different phases to assess the temporal variation in water quality. The methodology adopted for the analysis included the use of standard laboratory procedures to measure the various parameters. Statistical tools such as correlation and regression analyses were employed to identify the parameters that significantly affected the electrical conductivity of the water. Linear regression equations were formulated to determine the percentage contribution of each parameter to the EC. The results of the analysis showed that the water quality parameters varied across the different sampling sites and phases. The colour of the water ranged from 0.78 to 7 Hazen, with acceptable values for drinking purposes. The electrical conductivity values ranged from 123.7 to 472.5 µs/cm, within the prescribed limits by the Nepal Drinking Water Quality Standards. The study indicated that the presence of domestic and municipal wastes significantly influenced the water quality parameters. The analysis of total dissolved solids revealed values ranging from 59.95 to 256.5 mg/L, with higher values observed in areas influenced by domestic and municipal waste sources. Similarly, chloride content ranged from 9 to 42.99 mg/L, indicating relatively low contamination levels in the water samples. The study concluded that the hydro chemical analysis of the Ratuwa River and its tributaries provided valuable information for assessing water quality and identifying potential contaminants. The findings contributed to the understanding of the pollution status in the study area, particularly in the context of future industrial developments. The research highlighted the importance of regular monitoring and effective water treatment processes to maintain the quality of river water resources.

Mohammad Mirzavand et al. [50] focused on predicting groundwater level fluctuations in an arid environment. The study area was the Kashan Plain in Isfahan province, Iran. The researchers addressed the issue of decreasing groundwater levels in arid regions due to increasing water demand, weak irrigation management, and soil damage, which necessitated effective management and prediction of groundwater fluctuations. They utilized 36 piezometric wells and applied the ward algorithm to cluster the water table depths into five clusters. For each cluster, they employed five time series models: Autoregressive (AR), MA, ARMA, and SARIMA. The study found that the AR model with a two-times lag provided the most accurate groundwater level forecasting for a 60-month period ahead in all five clusters, with R2 score values ranging from 0.75 to 0.95. The results indicated that the average groundwater level fluctuation in 2010 and 2016 was 74.58 m and 80.71 m, respectively. Based on these conditions, the researchers estimated the groundwater depletion rate in 2016 to be 1.02

m per year. A statistical cluster analysis was conducted utilizing the ward algorithm to cluster the time series data of groundwater levels. The normalized groundwater level time series were analysed, and the resulting clusters were used to assign piezometers to their respective clusters. The results demonstrated that the AR (2) model exhibited the best performance in all five clusters, providing the highest accuracy in groundwater level forecasting. The effectiveness of combining multiple time series models was showcased, particularly the AR (2) model, in forecasting groundwater level fluctuations in an arid environment. The research provided valuable insights into the application of time series methods for groundwater forecasting and emphasized the significance of considering groundwater dynamics in arid environments.

Xuedi Zhang et al. [51] used geo-statistical methods and multivariate statistical analysis to examine the chemical characteristics of water samples collected from 39 sampling stations prior to the 2011 summer season irrigation period. The hydrogeological study conducted in the Yinchuan region of northwest China, the quality of phreatic water supplies in a semi-arid, traditional agricultural area was assessed. The objective was to identify the factors influencing phreatic water chemistry and measure its current status for effective water management. Through the application of principal component analysis (PCA) and cluster analysis, the study successfully identified the key factors that influence groundwater composition. The PCA revealed that the major variables influencing water quality in the study area were the evaporation effect caused by the dry climate, the dissolution of carbonate minerals containing F- and K-, and human activities such as domestic sewage treatment and chemical fertilization. Cluster analysis identified three distinct water types based on their chemical compositions and two clusters of sampling stations influenced by different sets of natural and/or anthropogenic factors. The dataset included measurements of major ions, total dissolved solids, total hardness, and other parameters following standard procedures recommended by the Chinese Ministry of Water Resources. The work results indicated that phreatic water in the Yinchuan region had weak alkaline properties and variable concentrations of major ions, with most samples exceeding China's acceptable limits for groundwater quality. The spatial distribution of hydro chemical variables exhibited variations across the study area, with higher concentrations of certain ions in specific regions. Correlation analysis identified significant relationships among the variables, providing insights into the hydro chemical processes in the groundwater system. PCA results showed that four principal components explained 87.6% of the total variance in the dataset. This study provided a comprehensive understanding of the hydrogeological characteristics and water quality pertaining to phreatic water supplies in the Yinchuan region. Through the adept application of multivariate statistical analysis techniques, the study

successfully identified the key factors that exerted influence over water chemistry and facilitated the classification of distinct water groups based on their chemical compositions. The implications of these findings extend to the realm of water management in the region, highlighting the imperative of considering both natural and anthropogenic factors when assessing groundwater quality.

2.2 TRADITIONAL MACHINE LEARNING APPROACHES IN WQI PREDICTION

Machine learning (ML) approaches have become increasingly popular for water quality index prediction due to their ability to handle large and complex datasets, capture non-linear relationships between variables, and adapt to changing conditions. Common ML algorithms used for water quality index prediction include random forests, SVMs, and ANNs. Random forests (RF) are useful for feature selection and identifying important predictors of water quality indices. SVMs and ANNs are capable of capturing complex relationships between variables. ML models have shown promising results in predicting water quality indices, but their performance was affected by overfitting model selection and data pre-processing. Therefore, it is important to carefully select the appropriate ML algorithm, optimize its parameters, and validate its performance on a range of datasets. Some of the traditional machine learning techniques that have been used previously are reviewed and stated below.

Semko et al. [52] focused on the development of precise and simple models for predicting water table fluctuations in the Lailakh Plain. ANN and Adaptive Neuro Fuzzy Inference Systems (ANFIS) were utilized as tools for modelling the nonlinear mappings in groundwater resource behaviour. Monthly average groundwater level, rainfall, temperature, and evaporation data were used to develop the proposed models. The data used in the models were normalized, and the training, validation, and testing datasets were determined. For training the neural network models, employed the multi-layer perceptron with the backpropagation algorithm. The study compared and prioritized the ANN and ANFIS models, including dynamic, static, and hybrid variants, using the Analytical Hierarchy Process (AHP). The ANN dynamic model, incorporating three input parameters, achieved the highest accuracy with a Mean Squared Error of 0.776 and a correlation coefficient of 0.975. The study underscored the significance of accurate groundwater level predictions for integrated management planning of groundwater resources. The research findings indicated that the dynamic and static models exhibited the highest accuracy for predicting groundwater fluctuations. Demonstrated the effectiveness of ANN and ANFIS models in predicting groundwater level fluctuations in the Lailakh Plain. The dynamic ANN model with three input

parameters was identified as the most accurate model for predicting water table fluctuations. These findings highlight the importance of accurate predictions in groundwater management and the sustainable use of water resources in the plain.

Batool et al. [53] investigated the quality and safety of spring water used for drinking purposes in Margalla Hills, Islamabad. The objective of the study was to assess the physicochemical and microbiological parameters of the spring water and determine if it posed any health risks to consumers. Fifteen water samples were collected from five different sites in Margalla Hills and analysed for various parameters, including colour, temperature, pH, odour, turbidity, hardness, total dissolved solids (TDS), EC, alkalinity, DO, chlorides, nitrates, sulphates, heavy metals, and microbiological indicators. A sampling approach was employed, collecting water samples from different sites in Margalla Hills, and both microbiological and physicochemical analyses were conducted using standard methods. Descriptive statistics and one-way analysis of variance were used for data analysis. The results of the study indicated that the spring water in Margalla Hills exceeded the World Health Organization (WHO) drinking water standards for EC, DO, cadmium, lead, and some microbiological parameters. The findings of the study indicated the presence of total coliform, Staphylococcus aureus, Enterococcus, and Pseudomonas aeruginosa in the spring water samples. The levels of total coliform were higher than the WHO standard, suggesting faecal contamination. The study also found variations in physical and chemical parameters such as pH, temperature, EC, alkalinity, and concentrations of nitrate, sulphate, sodium, potassium, and heavy metals among the different sampling sites. This research provided valuable insights into the quality of spring water in Margalla Hills and its potential impact on human health. The findings underscored the importance of regular monitoring and improvement of water quality standards to ensure the provision of safe drinking water to the local population.

Sakaa et al. [54] conducted a study aimed at developing a model for forecasting the water quality index in the Saf-Saf River using water quality parameters. The objectives of the study were to determine the importance of different input variables and assess the spatial and temporal variation in water quality. The study utilized data from 35 surface water samples that were collected from the Saf-Saf River basin. The water quality parameters were analysed using the CCME calculator software to calculate the WQI. Descriptive statistics, correlation matrix analysis, and multivariate statistical techniques such as PCA and FA were employed to analyse the data. The performance of the models was evaluated based on efficiency criteria and goodness-of-fit measures. The results indicated that the Multi-layer Perceptron (MLP) model exhibited the best performance with a small root mean square error (RMSE) value of 0.007 and

a high coefficient of determination of R2 score value was 0.811 compared to the other MLP models. The findings of the study highlighted the effectiveness of combining MLP neural networks and multivariate methods for assessing and forecasting water quality. The developed model and analysis techniques provide valuable tools for decision-makers to implement sustainable management practices. The MLP-based model demonstrated superior performance, and the analysis revealed spatial and seasonal variations in water quality.

Setshedi et al. [55] aimed to develop models using ANNs to predict water quality parameters in South Africa. The study utilized MLP and radial basis function (RBF) neural networks and collected data from three district municipalities in the Eastern Cape Province. Water samples were collected from rivers and wastewater treatment plants in the study area, and eight physicochemical parameters were analysed using standard methods. A comprehensive methodology involving sample collection, physicochemical analysis, and the use of ANN models for prediction was employed. The MLP and RBF neural networks were evaluated using training, validation, and testing sets. The ANN models were trained and tested using the collected data, and their predictive performance was evaluated. Two input combination models, MLP-4-5-4 and MLP-4-9-4, were compared based on their accuracy in predicting water quality parameters. The results showed that the MLP-4-5-4 model had a better understanding of the data and higher predictive ability compared to the MLP-4-9-4 model. The MLP-4-5-4 model achieved a correlation coefficient R2 score value of 0.989383 and a mean square error (MSE) of 39.06589 for the observed and predicted water quality. These findings have implications for natural water resources management in South Africa and similar catchment systems. This research provides valuable insights into the use of ANN models for predicting water quality parameters, contributing to the field of environmental monitoring and management.

Balraj Singh et al. [56] focused on the prediction of water quality using soft computing techniques. The objective of the study was to develop and compare the performance of three different soft computing techniques: ANN, Generalized Regression Neural Network (GRNN), and ANFIS for predicting the WQI in three sub-watersheds in Iran. The methodology employed in the study involved collecting flow and water quality data from Khorramabad, Biranshahr, and Alashtar sub-watersheds in Lorestan province, Iran. Ten physiochemical parameters were used as input variables, and the WQI, which represented water quality, was used as the output variable. The researchers applied ANN, GRNN, and ANFIS techniques to predict the WQI values. The results of the study indicated that ANN outperformed GRNN and ANFIS in predicting the WQI values. Among the different membership functions of ANFIS,

ANFIS_trimf showed better performance than the others. The findings suggested that ANN was a viable tool for predicting the WQI in water quality assessment. The study demonstrated the effectiveness of ANN and highlighted the advantages of using soft computing approaches in water resource management.

Monika Kulisz et al. [57] focused on the application of ANN methods for modelling the WQI in groundwater near shale gas extraction sites. Water samples were collected from 19 wells located in close proximity to a drilling site in eastern Poland. The objective of the study was to optimize the ANN by selecting the most relevant input parameters for predicting the WQI. The methodology involved collecting water samples from wells near a shale gas drilling rig and conducting standard laboratory analyses for various physicochemical parameters. The ANN models were developed using the RStudio and MATLAB environments, utilizing the Levenberg–Marquardt algorithm for training. The selection of input parameters was based on multiple regression analysis. Additionally, the hidden layer of the ANN consisted of five neurons. The results of the study provided insights into the groundwater quality in the vicinity of the drilling site, with a specific focus on the calculated WQI. The performance of this approach achieved satisfactory accuracy in predicting the WQI, with an RMSE of 0.651258, R-value of 0.9992, and R2 score value of 0.9984. The study demonstrated the effectiveness of ANN methods in predicting groundwater quality and offered an alternative to traditional WQI calculation methods. The utilization of advanced artificial intelligence can assist in water treatment and management, particularly in industrial areas. The research was considered novel as it applied ANN modelling to forecast WQI in an unstudied area in eastern Poland. The study also highlighted the importance of optimizing the number of parameters and the quality of the prediction models.

A. Solanki, et.al.[58], focused on predicting water quality parameters using deep learning techniques in the research. The objective of the study was to provide accurate predictions for variable data related to water quality, with a specific emphasis on addressing the issue of water pollution and its impact on human health and aquatic life. The research utilized secondary data collected from the Chaskaman River located near Nasik, Maharashtra, India. The analysis and prediction modelling were performed using the WEKA tool. The deep learning models employed included denoising autoencoder and deep belief network, known for their ability to handle data variability and provide accurate results. The research methodology involved data collection, data pre-processing to enhance data quality, and the development of predictive models using deep learning algorithms. Clustering techniques were applied to categorize the data based on seasons such as winter, summer, and monsoon, and missing values

were replaced with the mean of available values to clean the data. The results revealed that turbidity exhibited higher variation compared to pH and DO, with the most significant impact observed during the monsoon season. Conversely, pH remained relatively stable, while dissolved oxygen showed slight variations during summer due to temperature effects. The research highlighted the efficacy of deep learning-based prediction models in handling variable data and providing accurate predictions for water quality parameters. The findings suggested that the developed models were implemented for continuous monitoring of water quality, particularly in uncertain conditions.

Ahmed et al. [59] explored and evaluated an alternative method based on supervised machine learning for efficient prediction of water quality in real time. The study aimed to estimate the WQI and Water Quality Class (WQC) using a minimal number of input parameters and validated the feasibility of using this methodology in real-time water quality detection systems. The researchers utilized a dataset obtained from the Rawal watershed in Pakistan, collected by the Pakistan Council of Research in Water Resources. The dataset comprised 663 samples from 13 different sources of Rawal Water Lake, collected between 2009 and 2012. A series of supervised machine learning algorithms were employed to predict WQI and WQC based on four input parameters: temperature, turbidity, pH, and total dissolved solids. The methodology involved data preprocessing steps such as cleaning, normalization, and feature selection. Correlation analysis using the Pearson correlation method was performed to identify the relationships between the parameters. For regression algorithms, gradient boosting and polynomial regression yielded the most efficient predictions of WQI, with mean absolute errors (MAE) of 1.9642 and 2.7273, respectively. In terms of classification algorithms, MLP achieved the highest accuracy of 0.8507 in predicting WQC. The study compared its findings with previous research and highlighted the advantage of using a minimal number of parameters. Other studies typically employed more than 10 parameters, which were not suitable for inexpensive real-time systems. The methodology proposed here provided reasonable accuracy while using only four parameters, making it a viable option for real-time water quality detection. The findings laid the groundwork for developing an inexpensive real-time water quality detection system.

Sillberg et al. [60] have developed a machine learning-based technique integrating attribute-realization (AR) and SVM to classify the Chao Phraya River water quality. Using the linear function, the AR has identified the most important elements for improving river quality. NH3-N, Total Coliform (TC), Faecal Coliform (FC), Biological Oxygen Demand (BOD), DO, and Sal were the most contributing characteristics in the categorization, with contributed values

in the range of 0.80-0.98, compared to 0.25-0.64 for Total Dissolved Solids (TDS), Turb, Total Nitrogen (TN), NO3-N, and conductivity. The best classification results were achieved using the SVM linear approach, which had an accuracy of 0.94, a precision average of 0.84, a recall average of 0.84, and an F1-score average of 0.84. When applied to three to six parameters, the validation revealed that AR-SVM was a powerful method for identifying river water quality with 0.86-0.95 accuracy.

While traditional machine learning approaches have shown great promise for water quality index prediction, there are several potential disadvantages to their use. One of the main concerns is overfitting, which can occur when a model is overly complex and is trained too closely to the training data. This can lead to poor performance when the model is applied to new data. Additionally, traditional ML models require large amounts of high-quality data, which may not always be available or easy to obtain. There is also a risk of bias if the training data is not representative of the target population or if there are errors or missing values in the data. Therefore, while ML approaches hold great promise for water quality index prediction, it is important to carefully consider the limitations and potential pitfalls of these methods.

2.3 DEEP LEARNING APPROACHES IN WQI PREDICTION

Deep learning approaches have gained significant attention and demonstrated promising results in various domains, including environmental science and water quality assessment. In the context of river water quality prediction, deep learning techniques have the potential to enhance the complex interactions between environmental factors and water quality parameters. By analysing and synthesizing the existing research, this review aims to identify the state-of-the-art methodologies, challenges, and future directions in utilizing deep learning techniques for WQI prediction. Through this investigation, valuable insights have been gained to facilitate the development of accurate and efficient models for assessing and managing water quality in river systems. Some of the deep learning architecture-based research works are reviewed and stated below.

Wang et al. [61] proposed a water quality prediction method based on the long and short-term memory neural network (LSTM-NN). The study aims to address the limitations of traditional neural networks in handling time series data and improve the accuracy and generality of water quality prediction. The study aims to address the limitations of traditional neural networks in handling time series data and improve the accuracy and generality of water quality prediction. The researchers established a prediction model based on LSTM-NN for water quality prediction. They used a dataset of water quality indicators in Taihu Lake,

measured monthly from 2000 to 2006, as the training data. The proposed method was compared with two other methods: one based on a backpropagation neural network (BP-NN) and the other based on an online sequential extreme learning machine (OS-ELM). The results showed that the water quality prediction method based on LSTM-NN outperformed the other two methods in terms of accuracy and generality. The authors specifically focused on predicting DO and TP values in Taihu Lake and demonstrated the effectiveness of the approach. The research article presented a water quality prediction method based on LSTM-NN, which addresses the challenges of handling time series data. The study demonstrated that the proposed method offers improved accuracy and generality compared to traditional neural networks and other prediction methods. The findings highlighted the potential of deep learning techniques, specifically LSTM NN, in water quality prediction.

Zhenbo et al. [62] proposed a hybrid model that combined a sparse auto-encoder (SAE) and LSTM for the purpose of improving the prediction accuracy of dissolved oxygen in aquaculture and water quality prediction. The SAE was employed to pre-train the hidden layer data and capture deep latent features related to water quality, which were subsequently fed into the LSTM to enhance the accuracy of predictions. The SAE, serving as a feature extraction pre-training network, effectively improved the performance of LSTM and BPNN models in predicting water quality factors. The authors suggested that deep learning techniques, such as autoencoders and LSTM, offered distinct advantages in terms of feature extraction and the processing of sequence data, thereby offering new perspectives for research in water quality prediction. Experimental results demonstrated that the proposed SAE-LSTM model consistently outperformed the LSTM model, achieving reductions in MSE of 23.3%, 53.6%, and 39.2% for prediction steps of 3, 6, and 12 hours, respectively. Moreover, the SAE-LSTM model exhibited superior performance compared to the SAE-BPNN model, achieving MSE reductions of 87.7%, 91.9%, and 90.0%. Notably, the prediction accuracy of the SAE-LSTM model decreased with increasing prediction steps. Although the combined SAE-LSTM model exhibited slightly slower performance than the single LSTM model, it yielded smaller errors. Additionally, the MSE reductions achieved by the SAE-BPNN model indicated improvements over the standard BPNN model. These findings underscore the potential of deep learning methods and indicate potential avenues for future research aimed at enhancing prediction accuracy in the field of aquaculture water quality.

Jianlong Xu et al. [63] proposed a time series prediction method, with seq2seq framework for accurate and efficient water quality prediction. Addressed the challenges posed by inconsistent data acquisition frequency, data organization, and the volatility and sparsity of

water quality data. The approach employed in this study was based on a sequence-to-sequence (seq2seq) framework, where the GRU model served as both the encoder and decoder. Additionally, a factorization machine (FM) was integrated into the model to effectively handle the high sparsity and high-dimensional feature interactions in the data. In an effort to overcome the memory constraints often experienced in deep learning when handling large amounts of temporal water quality data, a dual attention mechanism was incorporated into the seq2seq framework. The research conducted experiments utilizing water quality data obtained from the Lianjiang River in Guangdong, China. The results demonstrated that the proposed FM-GRU method outperformed several typical water quality prediction methods in terms of accuracy. The findings of the study included the successful application of factorization machines to time series forecasting, enabling the extraction of high-dimensional feature information without the necessity of traditional artificial feature engineering. Based on the research, it was concluded that the FM-GRU model represents a reliable and effective method for water quality prediction. By leveraging the strengths of the FM model for feature extraction and the seq2seq framework for encoding and decoding, the FM-GRU model provided accurate predictions.

Yilma et al. [64] employed an ANN to simulate the WQI and addressed the issue of water pollution in the Little Akaki River, Ethiopia. The objective of the study was to gain a comprehensive understanding of the pollution status of the river water and provide support for decision-making in water resource management. Water samples were collected from 27 sites during the dry and wet seasons in 2015 and 2017 for analysis. Twelve water quality parameters, including pH, temperature, total suspended solids (TSS), TN, BOD, DO, total organic carbon (TOC), and chemical oxygen demand (COD), were considered in determining the WQI. Descriptive statistics were utilized to analyse the data and evaluate the river water quality condition. The findings revealed generally poor water quality in the Little Akaki River, with high levels of TSS, T-N, and BOD, and low levels of DO. Spatial variations in water quality were observed, with higher pollution levels found in densely populated areas. The calculated WQI values classified the river as impaired, except for the headwater site, which exhibited fair water quality. To predict the WQI, an ANN model was trained and validated using various combinations of hidden layers, neurons, and activation functions. The optimized ANN model, consisting of eight hidden layers, 15 hidden neurons, logsig activation function, and purelin output function, demonstrated a high level of agreement between the calculated and predicted WQI with an R2 score value of 0.93. This successful application of the ANN model showcased its effectiveness in modelling the water quality index of the Little Akaki River. Furthermore,

the study emphasized the significance of BOD as a crucial water quality parameter and its relationship with the WQI.

Abhay Srivastava [65], analysed and forecasted the pH levels of rivers and their influence on aquatic ecosystems was also investigated. The objective of the study was to analyse and forecast river pH levels using deep learning techniques and assess the importance of temperature as a predictive feature. Various deep learning approaches, including LSTM, GRU, RNN and TFT models, were compared to determine the most effective algorithm for pH prediction. The methodology encompassed several steps, beginning with the collection and cleansing of Virginia Current Conditions Water Quality data, with the exclusion of irrelevant parameters. The dataset, represented as a time series, included temperature as a pertinent factor impacting pH levels. Missing values were addressed through time-based interpolation, and the data was resampled on a daily basis to reduce granularity. Exploratory data analysis was conducted, and in the data preprocessing phase, temperature and time-based variables were incorporated to account for seasonal variations in pH. Different normalization techniques were applied to the TFT, LSTM, GRU, and RNN models. The study also employed hyperparameter tuning using Optuna to optimize the TFT model. The best trial from the optimization process was selected based on validation loss, and the corresponding hyperparameters were utilized for the final TFT model. Model performance was assessed using metrics such RMSE. The findings of the study revealed that the TFT model outperformed other deep learning methods in pH level prediction. Temperature emerged as a significant feature in pH prediction, highlighting the potential impact of global warming on pH levels. The study successfully predicted pH anomalies and identified significant disparities between the original and predicted data for nine out of the ten analysed rivers. The TFT model exhibited accuracy, with an average RMSE of 0.217 on the validation datasets. Overall, this research article contributed valuable insights into the understanding of pH levels in rivers and presents a deep learning-based approach to pH analysis and forecasting. The results highlighted the importance of temperature and demonstrated the superiority of the TFT model in predicting river pH levels.

A comprehensive review of the existing literature reveals that numerous researchers have conducted analyses on the quality and quantity of various water resources worldwide. The details of literature review are summarized in Table II. These studies primarily focused on assessing parameters such as pH, TDS, EC, sodium, potassium, calcium, magnesium, fluoride, and chloride, among others. Furthermore, researchers have extensively explored the concept of the WQI and employed diverse machine learning algorithms for building prediction models.

Table II. Summary of Literature Review

Authors	Objective	Methodology	Data	Prediction Rate	Findings
Shrestha, AK & Basnet, N. 2018,	To analyse and interpret the hydro chemical parameters of the river water and evaluate the percentage contribution to the electrical conductivity	Correlation and regression analysis	Two different phases of Ratwa River	Conductivity values :123.7 - 472.5 µs/cm	Analyses stated that parameters are above the limit
Mohammad Mirzavand & Reza Ghazavi	To forecast groundwater level in the environment using time series methods	Autoregressive, MA, ARIMA, SARIMA	Groundwater level for a 60-month	ARIMA-R2 Score: 0.75 - 0.95	ARIMA model yielded high efficiency
Semko Rashid,Milad Mohammadan, koorosh Azizi	To predict the groundwater level fluctuation using ANN and ANFIS in Lailakh Plain	ANN and ANFIS	Rainfall, temperature, evaporation	MSE :0.776 R2 Score : 0.975	Dynamic ANN outperformed
Xuedi Zhang, Hui Qian, Jie Chen and Liang Qiao	To assess the groundwater chemistry and status in a heavily used semi-arid region with multivariate statistical analysis	Principal Component Analysis	39 sampling stations of Yinchuan region	total variance: 87.6%	Analysed the ground water status using PCA
Batool A., Samad N., Kazmi S. S., Ghufran M. A., Imad S., Shafqat M. and Mahmood T.	To assess the physicochemical and microbiological parameters of the spring water	ANOVA	Temperature, turbidity, hardness, TDS, EC, DO, and others	Limits are varying for parameters	physico-chemical parameters were found below the WHO recommended limit
Sakaa, B., Brahmia, N., Chaffai, H. and Hani, Desalin.	To determine the importance of different input variables and assess the spatial and temporal variation in water quality	Neural networks and multivariate methods	Physico-chemical Parameters from Saf-Saf River	MLP- R2 Score: 0.811	MLP yielded an R2 Score as compared to other models
Setshedi, K. J., Mutingwende, N. and Ngqwala,	To develop models using Artificial Neural Networks for predicting water quality parameters	MLP Types	Tyhume, Bloukrans, Buffalo Rivers, wastewater treatment plants	MLP-4-9-4: R2 Score - 0.98	MLP-4-9-4 model R2 Score as compared to other prediction models
Singh, B., Sihag, P., Singh, V. P., Sepahvand, A. and Singh, K.	To predict the groundwater level fluctuation using ANN and ANFIS in Lailakh Plain	Soft Computing Techniques	Khorramabad, Biranshahr, and Alashtar sub-watersheds in Iran	ANN - R2 score: 0.99	ANN technique outperformed both the GRNN and ANFIS techniques
Kulisz, M., Kujawska, J., Przysucha, B. and Cel, W	To forecast the water quality index in groundwater using artificial neural network	Neural network modelling	Physiochemical Parameters from Warta River	R2: 0.9984	ANN model performer better than in predicting WQI

A. Solanki, H. Aggarwal, and K. Khare	To predict the water quality parameters using Deep Learning	Denoising auto-encoder, deep belief networks, MLP	pH, turbidity, dissolved oxygen from Krishna River basin	DBN- R2 score 0.89	DBN achieved high R2 score as compared to other prediction models
U. Ahmed, R. Mumtaz, H. Anwar, A.A. Shah, R. Irfan, J. Garca-Nieto	To predict t water quality efficiently with supervised machine learning algorithms	Gradient boosting and MLP	Rawal water shed, situated in Pakistan	MLP-accuracy: 0.85	MLP yielded high accuracy as compared to other models
C.V. Sillberg, P. Kullavanijaya, O. Chavalparit	To develop a machine learning-based technique integrating attribute-realization and support vector machine to classify the Chao Phraya River water quality	Attribute-realization and support vector machine	NH3-N, TCB, FCB, BOD, DO	AR-SVM-accuracy: 0.95	Integrated model AR-SVM performs better than other models
Wang, Y., Zhou, J., Chen, K., Wang, Y., & Liu, L.	To predict water quality index using LSTM neural network	LSTM and neural network	Taihu Lake, measured monthly from 2000 to 2006	LSTM RMSE: 0.046	LSTM predicts WQI with less error rate
Zhenbo Li, Fang Peng, Bingshan Niu, Guangyao Li, Jing Wu, Zheng Miao	To predict water quality by combining sparse auto-encoder and LSTM network	Sparse Auto-encoder and LSTM Network	DO, temperature, ammonia, pH, humidity, wind speed	SAE-LSTM accuracy: 91.9	Integrated SAE-LSTM achieved better results than other models
Xu J, Wang K, Lin C, Xiao L, Huang X, Zhang Y.	To predict time Series water quality based on seq2seq framework	FC- LSTM, FC-GRU, FM-GRU	water quality data from Lianjiang River	FM-GRU – RMSE: 0.4	FM-GRU performed better with less error rate
M. Yilma, Z. Kiflie, A. Windsperger, N. Gessese	To predict WQI using artificial neural network: a case study in little Akaki River	ANN	Akaki River Monitoring Stations	ANN - R2 score: 0.93	ANN yielded a high R2 score value
Srivastava, A., Cano, A.	To analyse and forecast the rivers pH level using deep learning.	RNN, GRU, LSTM and TFT	PH data from a river basin	TFT-RMSE: 0.217	TFT yields better accuracy

In the existing research, several aspects have not been adequately addressed in the field of water quality prediction models. Firstly, there is a lack of integration of heterogeneous data sources such as climate data. Incorporating additional data sources could offer a more comprehensive understanding of water quality dynamics. Secondly, real-time prediction, which involves incorporating up-to-date data and providing immediate forecasts, has not been extensively explored compared to the predominant focus on offline prediction using historical data. Lastly, the inclusion of uncertainty analysis in water quality prediction models remains limited, despite its importance in achieving reliable predictions. Addressing uncertainty

through estimation and propagation would enhance the reliability and applicability of these models.

Many studies have limited spatial and temporal coverage, focusing on specific water bodies or restricted geographical areas, thereby limiting the generalizability of the models and inadequately capturing seasonal and interannual variations. The insufficient consideration of non-linear relationships in widely used linear regression models hinders their ability to capture complex relationships between water quality parameters and influencing factors, necessitating the incorporation of more advanced non-linear models. The lack of interpretability in some machine learning and deep learning models poses challenges in understanding the underlying factors contributing to water quality changes, emphasizing the need for interpretable models to provide insights into important features and causal relationships.

However, a research gap exists in the development of prediction models for the water quality index specifically for the Bhavani River and Bharathapuzha River in South India. Additionally, there is a scarcity of literature addressing the water quality prediction models specifically tailored for the South Indian River systems. In light of this knowledge gap, the current study aims to investigate the water quality of the Bhavani River and Bharathapuzha River. Informed by the existing literature, advanced deep learning modelling techniques have been recognized as efficient tools for WQI prediction. Thus, this study employs hybrid approaches to develop predictive models for the river water quality index.

Remarks

The paper entitled "Predictive Models for River Water Quality using Machine Learning and Big data Techniques - A Survey" has been presented in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, March 25-27,2021, and published in IEEE Digital Xplore, pp. 1747-1753. ***(Scopus Indexed)***