J. Ind. Geophys. Union, 30(1) (2026), 70-77
https://doi.org/10.71122/JIGU.30(1)2026.007

Deep learning-based model for groundwater quality prediction in Kanyakumari District, Tamil Nadu, India.

Annie Jose* and Srinivas Y
Centre for Geotechnology, Manonmaniam Sundaranar University, Tirunelveli – 627 012, Tamil Nadu (India)
*Corresponding Author: anniejose1996@gmail.com

ABSTRACT

In this study, a Deep learning-based model, Stochastic Neural Network (SNN), which excels in handling uncertain and complex data, is customized to predict groundwater quality in the Kanyakumari District of Tamil Nadu (India), where groundwater is considered a critical source for drinking and agricultural purposes. The SNN model captures the stochastic nature of the data and provides reliable predictions by simulating multiple possible outcomes, making the model ideal for groundwater quality prediction. A range of groundwater quality indicators, such as pH, EC, TDS, etc. were used to train the model. Rainfall patterns were found to have a considerable impact on water quality over a ten-year period, emphasizing the importance to include seasonal data into prediction models. Hence, rainfall data was also included in order to evaluate its impact on groundwater quality. The deep learning model demonstrated its effectiveness with 95% prediction accuracy. The model's capacity to distinguish between classes was evaluated by the Classification Report, Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC) values and the confusion matrix. In addition to this, Cross-validation (CV) was employed to confirm the model's performance and also to test the reliability of the results. This study provides an efficient method that can assist in sustainable use of groundwater resources.

Key Words: Deep learning, Groundwater quality prediction, Rainfall, Stochastic Neural Network, Kanyakumari District, Public Health.

INTRODUCTION

Water is a necessity for survival, economic development and human well-being, hence access to clean water is a fundamental right (Kirsch, 2006). With rising living standards, population demands, and expanding domestic, industrial, agricultural, and urban activities, water consumption has increased, resulting in overuse and deterioration of both surface and groundwater resources (Gilli et al., 2012). Groundwater in particular, is difficult to manage and more susceptible to contamination from both geogenic and anthropogenic sources (Karamouz et al., 2011). As water easily dissolves substances, monitoring groundwater quality changes is critical for its protection and conservation. Exceeding permissible limits of physicochemical parameters, may pose serious health risks (Karnena and Saritha, 2019). Groundwater is vital for drinking as well irrigation, especially in the study area of Kanyakumari, in Tamil Nadu and hence proper quality monitoring is critical for sustainable resource management. Given the complex and variable nature of groundwater, which is influenced by both natural and human influences (Giao et al., 2023), traditional evaluation methods based on laboratory testing and statistical analysis can be time and resource draining, especially for large datasets. To address these limitations, this study utilizes data-driven modelling to improve the accuracy and reliability of groundwater quality predictions.

Over the last decade, Machine Learning (ML) has grown rapidly in groundwater quality modeling, as ML algorithms learn patterns directly from data rather than depending on pre-set equations, allowing for precise predictions of water quality parameters (Haggerty et al., 2023). Deep Learning (DL), a subset of machine learning, inspired by neural processing in the human brain, employs multilayer neural networks to evaluate complicated, structured data (Chollet, 2017). These developments indicate great potential in groundwater research, as neural networks often deliver accurate and dependable predictions. They have been extensively used to predict groundwater levels, evaluate water quality near shale gas reserves, improve MODFLOW simulations, predict nitrate contamination and model groundwater drawdown (Lohani and Krishnan, 2015; Kulisz et al., 2021; Gholami and Sahour, 2022; Stylianoudaki et al., 2022; Kishor et al., 2025). All studies indicate that the neural networks may effectively assess groundwater quality and quantity, but most studies utilize conventional Artificial Neural Networks (ANN), which do not capture or forecast uncertainty. Stochastic Neural Networks (SNNs), on the other hand, can reduce overfitting and capture uncertainty. Despite being widely used in domains such as, finance (Kalariya et al., 2022), civil engineering (Emig et al., 2023), microbiology (Sarmadi et al., 2022), and physics (Schneider et al., 2017), their use in groundwater research remains limited. SNN's strength lies in making probabilistic predictions and adapting to changing environmental conditions. In this study, SNN with Monte Carlo Dropout, was used to simulate numerous possibilities per input, resulting in uncertainty-aware groundwater quality prediction. Rainfall was also included to better express climate-related effects on groundwater chemistry. The main objective of the study is to create a sophisticated deep learning model that is specifically suited to capturing the diversity and complexity present in groundwater quality data, to study the prediction reliability using stochastic inference and assess how rainfall affects the accuracy of water quality classification in the Kanyakumari District of Tamil Nadu, India.

STUDY AREA

The study area, Kanyakumari District (Figure 1), is one of those areas that heavily relies on groundwater as a freshwater supply for agriculture and drinking. The study area falls between 8º 03' to 8º 35 ' N and 77º 15' to 77º 36' E, which is located in the southernmost tip of India. It is bounded by three water bodies: Indian Ocean, Arabian Sea and the Bay of Bengal. The study area enjoys both northeast (NE) and southwest (SW) monsoon seasons, resulting in about receiving an approximate 1448.6 mm of precipitation on average, every year (District Statistical Handbook, 2021-2022). According to GSI (2005), groundwater is present in almost all the formations. The studied district is composed of 80% fractured hard rocks and 20% sedimentary terrain.

METHODOLOGY

Data preparation

The overall methodology of the study is shown in the Figure 2. A decade (2013-2022) worth of groundwater quality data, which comprises of 65 locations and rainfall data from 25 rain gauge stations, were collected from State Ground and Surface Water Resources Data Centre, "Public Works Department", Government of Tamil Nadu. Rainfall patterns are one of those parameters that has a greater influence in the groundwater quality. And hence, rainfall is also added along with groundwater quality parameters such as pH, electrical conductivity (EC), total dissolved solids (TDS), major cations such as Na^+, K^+, Ca^{2+}, Mg^{2+}, total hardness (TH, calculated as the sum of Ca^{2+} and Mg^{2+}), and anions such as SO_4^{2-}, Cl^-, HCO_3^-, CO_3^{2-}, F^-, NO_2^- +NO_3^-. Information such as location, name of the village and the date of collection were also taken along with the parameters. Over the course of the ten-year period, a total of 645 data points were obtained from groundwater samples taken from 65 monitoring wells, spread throughout the study region. Samples were taken twice a year, in the pre-monsoon and post-monsoon seasons, and each site provided several observations. Theissen Polygon Method, which allocates rainfall values to each sampling location according to physical proximity and area of effect, was used to integrate rainfall data with the groundwater sampling sites. The raw data were unsuitable for use without pre-processing and hence, it is carried out by handling missing values and any type of inconsistencies in the dataset was removed to maintain the quality of the data. Missing data and outliers were addressed using simple deletion to assure the accuracy of the analysis and only complete and validated records were included. Outlier detection was also aided by the Calculation of Ion Balance Error (IBE) and data points with notable charge imbalance excluded.

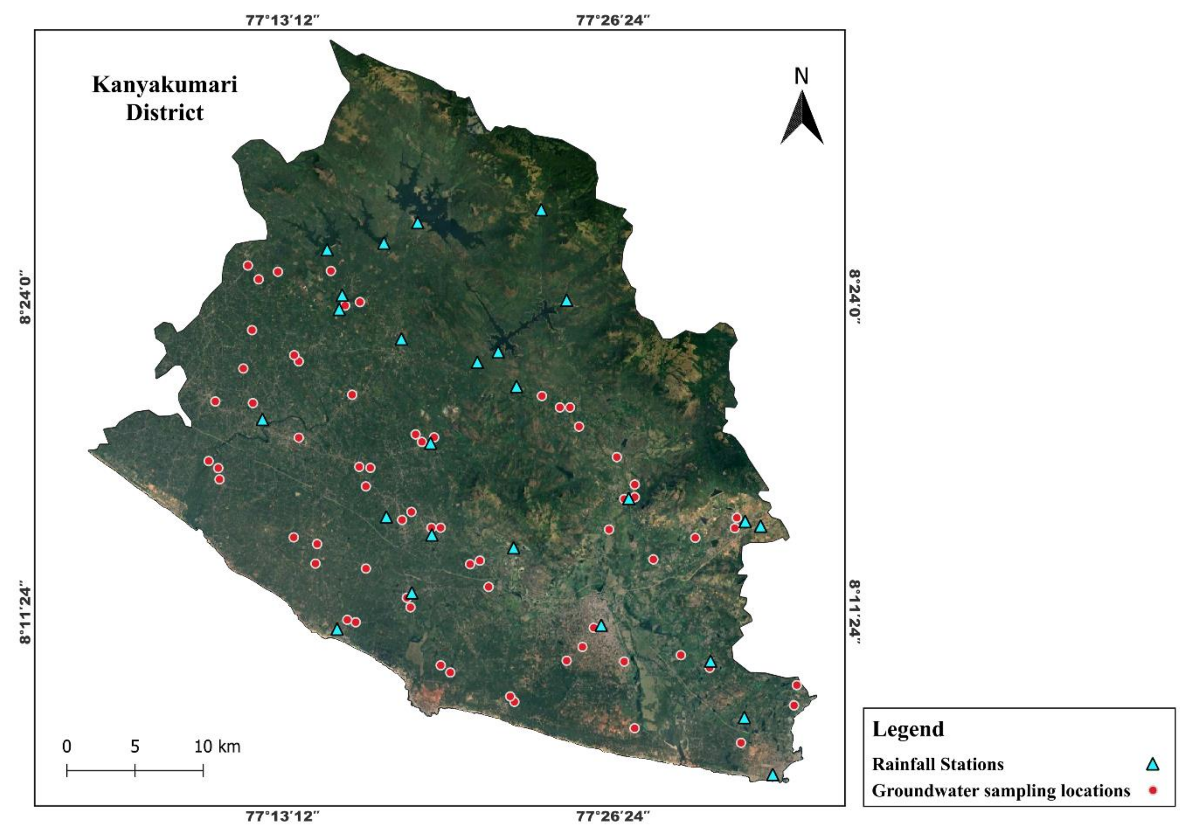

Figure 1. Map showing the distribution of sampling locations and rainfall monitoring stations across the study area in the Kanyakumari District, Tamil Nadu, India.

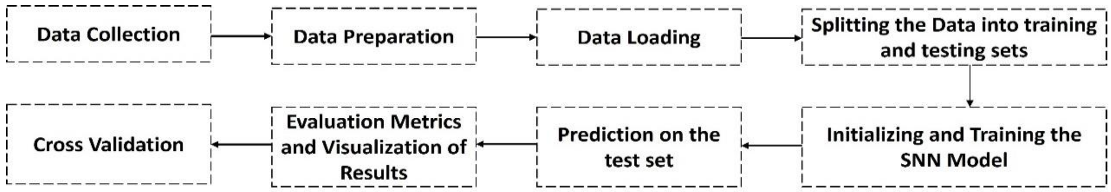

Figure 2. Methodological framework outlining the sequential processes involved in the study.

IBE also known as Charge Balance Error (CBE), is a quality control measure used to evaluate the reliability of analytical data in water chemistry analysis and also to remove any other inconsistencies in the dataset. The calculated IBE had to fall within the permissible range of ±10% (Domenico and Schwartz, 1990). Then, Water Quality Index (WQI) for drinking is calculated on the relation based on BIS (2012) and WHO (2011), which is followed in many studies for example, Vasanthavigar et al. (2010). The purpose of computing Drinking Water Quality Index (DWQI) is to help the researchers, policymakers and the general public to quickly assess the safety of drinking water and make knowledgeable choices regarding water treatment and consumption by condensing complex water quality data into a single numerical value based on multiple water quality parameters. It is expressed as,

$$DWQI = \sum_{i=1}^{n} SI\ i$$

Where 'SI' is the Sub- index which is calculated from assigned parameter weight, measured concentration (C_i) and the permissible limit (S_i) based on BIS (2012) and WHO (2011) standards and 'n' is the number of parameters.

After the calculation of Water Quality Index, the results obtained are compared with the Water Quality Index (WQI) classification in India, which categorizes groundwater into five quality levels, helping in determining the suitability of groundwater for consumption and indicating the severity of water quality degradation.

The classes are labeled as Class 0, Class 1, Class 2, to be used in the prediction. Following the DWQI, the correlation matrix (**Figure 3)** exhibits linear connections between groundwater quality metrics, with coefficients ranging from -1 to +1. Strong positive correlations were found between EC, TDS, Na, and Cl (r > 0.85), showing a strong hydrochemical relationship, whereas fluoride, and nitrate showed weaker relations.

Overall, the matrix highlights significant hydrochemical linkages and aids in feature selection for the deep learning-based groundwater quality prediction model.

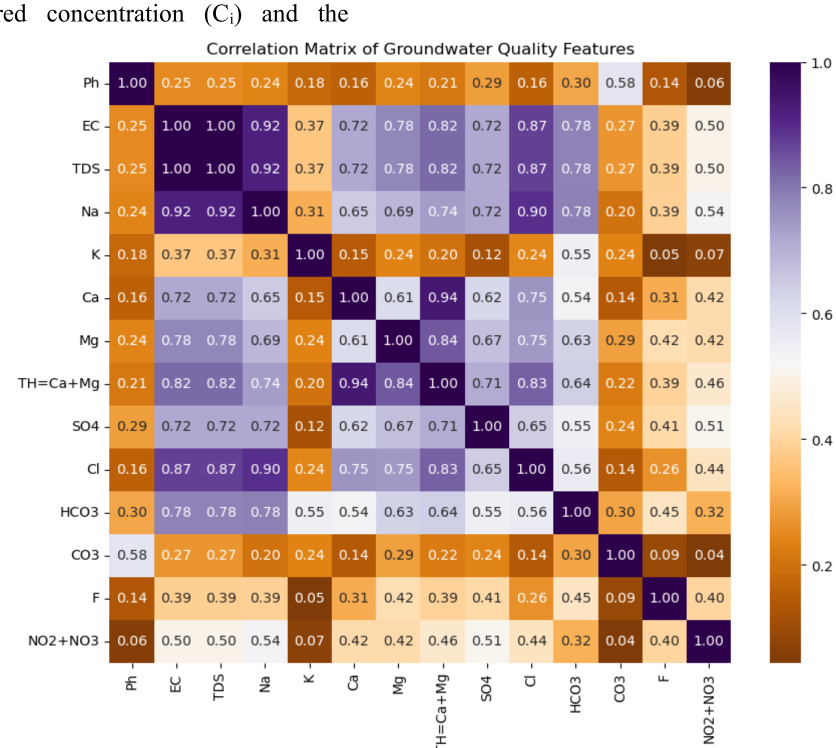

Figure 3. Pearson correlation matrix depicting the strength and direction of linear correlations between groundwater quality metrics, generated to assist in feature selection.

Working of Stochastic Neural Network (SNN) model

Based on Florensa et al. (2017), a broad category of neural networks incorporating stochastic units in the computation are called Stochastic neural networks (SNNs). The SNN model is customized based on the necessities of the research work. This study uses SNN with Monte Carlo Dropout, an effective deep learning method for classification problems that gives uncertainty estimation in addition to predictions. The Framework of SNN model for groundwater quality prediction involves the process of data normalization, assignment of weightages on the neurons, regularization and model compilation after data preparation. Data normalization process is done to ensure that the input features follow a typical Gaussian distribution, a fundamental requirement for accurate and reliable model training called Z-score normalization having been used to normalize the features.

The architecture of the SNN was designed with an input layer, two hidden layers and an output layer. The layout of the input layer is defined in accordance with the number of features. The first hidden layer has 128 neurons with ReLU (Rectified Linear Unit) Activation function and the second hidden layer has 64 neurons. The ReLU activation function was chosen for its computational simplicity, strong gradient propagation, and ability to prevent the vanishing gradient problem, which was observed in trials with deeper networks and in environmental datasets with wide range of parameters. The output layer consists of three units with Softmax activation function. After assigning the neurons, Dropout is used as a regularization method to reduce overfitting. Using Monte Carlo Dropout, the model may assess uncertainty during inference by producing a range of predictions rather than a single output, allowing it to better capture variability and prediction confidence. The model was trained on 80% of the data and tested on the remaining 20%, then combined with the Adam optimizer and Sparse Categorical Cross-Entropy loss. To reduce overfitting, a learning rate of 0.001, 10 epochs, and a batch size of 32, with a 0.5 dropout rate after each dense layer, were employed. Monte Carlo Dropout with 100 stochastic iterations was utilized to assess prediction uncertainty, and early stopping based on validation loss (patience of three epochs) guaranteed training was stopped once convergence was achieved.

Evaluation metrics

In order to evaluate the effectiveness of the deep learning model, computation of evaluation metrics is essential (Müller and Guido, 2016). To guarantee robustness and dependability, the performance of the suggested SNN model was thoroughly assessed using an array of statistical and graphical metrics. Metrics such as Confusion Matrix, Classification Report, Receiver Operating Characteristic (ROC) and Area Under Curve (AUC), are computed to assess the model. According to Pradhan and Kumar (2019), a confusion matrix, also referred to as an error matrix, is generated by comparing the observations on actual and predicted values of the dataset. For a multiclass classification problem, the confusion matrix is taken as 3×3 matrix for three classes, with elements - True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Next, Classification Report is generated which summarizes significant metrics such as accuracy (the percentage of correctly predicted samples), precision (the percentage of relevant instances among predicted positives), recall (the percentage of relevant instances correctly identified), and F1-score (the harmonic mean of precision and recall). ROC curve, a graphical plot illustrates the diagnostic ability of a classifier system as its discrimination threshold is varied and AUC is a numerical value that represents the area under the ROC curve. A higher AUC generally indicates a better-performing model. Cross-validation was applied to assess the model's generalization capabilities; with K-fold CV chosen for its reliability, especially when working with limited environmental datasets. The five-fold CV ensured that each sample contributed to both training and validation, which improved model stability and reduced bias. The model was trained on 80% of the data in each fold and validated on the remaining 20%, and the process was repeated five times to ensure a thorough performance evaluation.

RESULTS AND DISCUSSION

Groundwater quality in the study area was classified into three categories: Excellent (Class 0), Good (Class 1), and Poor (Class 2). Figure 4 reveals that 77% of samples are rated excellent, 18% good, and 5% poor, indicating an evident class imbalance. To address this issue, a class-weighting approach was utilized during model training, with weights assigned in inverse proportion to class frequencies in order to prioritize minority samples while reducing bias toward the majority class. Also, Figure 5 suggests that rainfall patterns play a crucial role in influencing the groundwater quality. As rainfall increases, water quality tends to improve and vice-versa.

Performance of the SNN model across three categories Class 0 (excellent water), Class 1 (good water), and Class 2 (poor water) are depicted in the confusion matrix (Figure 6). With only three samples incorrectly classified as Class 1, the model performs exceptionally well, particularly for Class 0. Class 2 instances are correctly predicted but the iterative instances are very small, may be due to the very low number of data points. The model is dependable for practical groundwater quality evaluation as evidenced by its low misclassification, especially when it comes to detecting good and poor water quality.

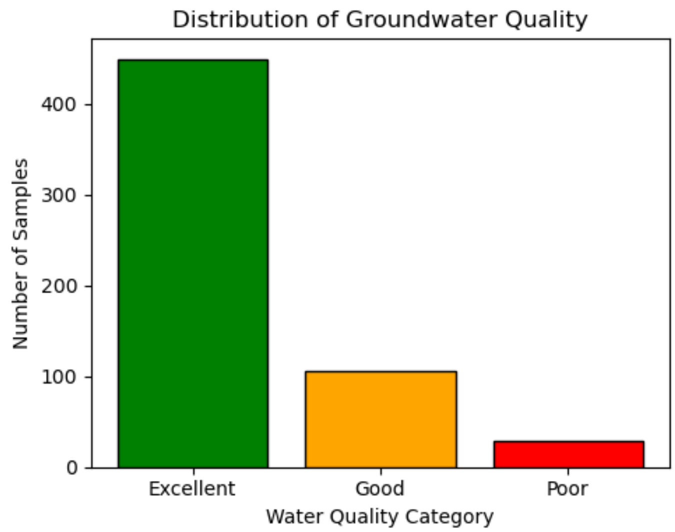

Figure 4. Bar chart illustrating the distribution of groundwater quality classes (excellent, good, and poor) across the study area, based on the Drinking Water Quality Index (DWQI).

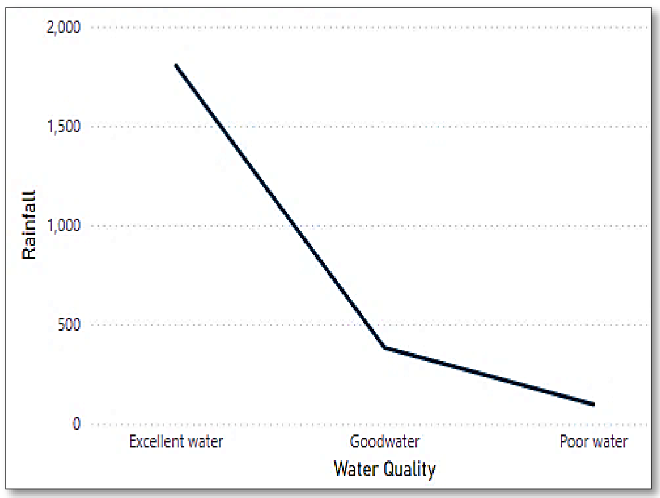

Figure 5. Graph depicting the relationship between groundwater quality classes and rainfall variation in the study area.

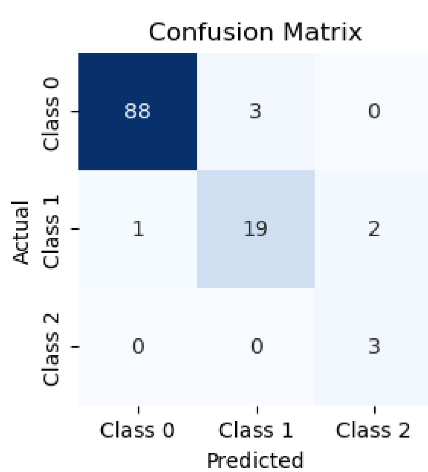

Figure 6. The confusion matrix displaying the classification performance of the Stochastic Neural Network (SNN) model across the three groundwater quality categories.

The classification report (Table 1) demonstrates impressive model performance, with Class 0 attaining near-perfect scores due to its more widespread inclusion in the dataset, Class 1 exhibiting solid precision and recall (0.86), and Class 2 with reduced accuracy but perfect recall despite limited data. Overall, the model reached 95% accuracy, indicating a strong and dependable groundwater quality classification.

To visualize the model's performance across various threshold settings, ROC curves (Figure 7) were generated, giving a graphical representation of the true positive rate (sensitivity) against the false positive rate. The Area Under the Curve (AUC) values for all classes are high, indicating strong classification performance. The ROC-AUC for each class were satisfactory, indicating strong discriminatory ability.

The prediction graph (Figure 8) displays a clear agreement between the actual groundwater quality classes and the SNN model's predicted labels, with real values represented by a blue solid line and forecasts by an orange dotted line. The close overlap of the two curves implies strong sample-wise prediction accuracy. The model's resilience is further confirmed by its high mean cross-validation accuracy of 93% and low standard deviation of 0.03, indicating consistent and reliable performance across all folds.

Table 1. Classification report on the SNN Model.

Class labels	Precision	Recall	F1 score	Support
0	0.99	0.97	0.98	91
1	0.86	0.86	0.86	22
2	0.60	1.00	0.75	3
Accuracy			0.95	116
Macro avg.	0.82	0.94	0.86	116
Weighted avg.	0.95	0.95	0.95	116

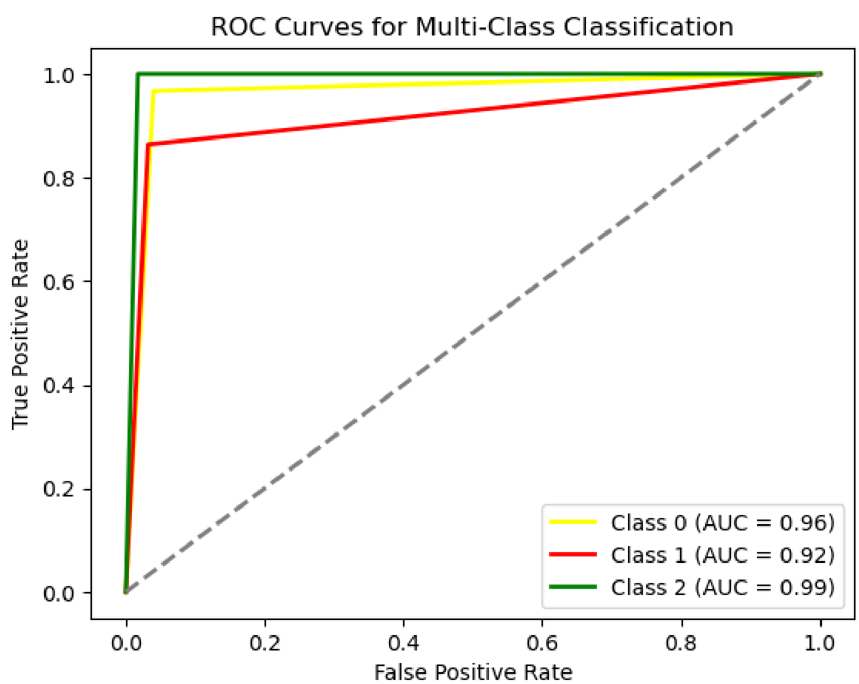

Figure 7. The smooth and well-separated Receiver Operating Characteristic (ROC) curves with corresponding Area Under the Curve (AUC) values for the Stochastic Neural Network (SNN) model generated for the three groundwater quality classes

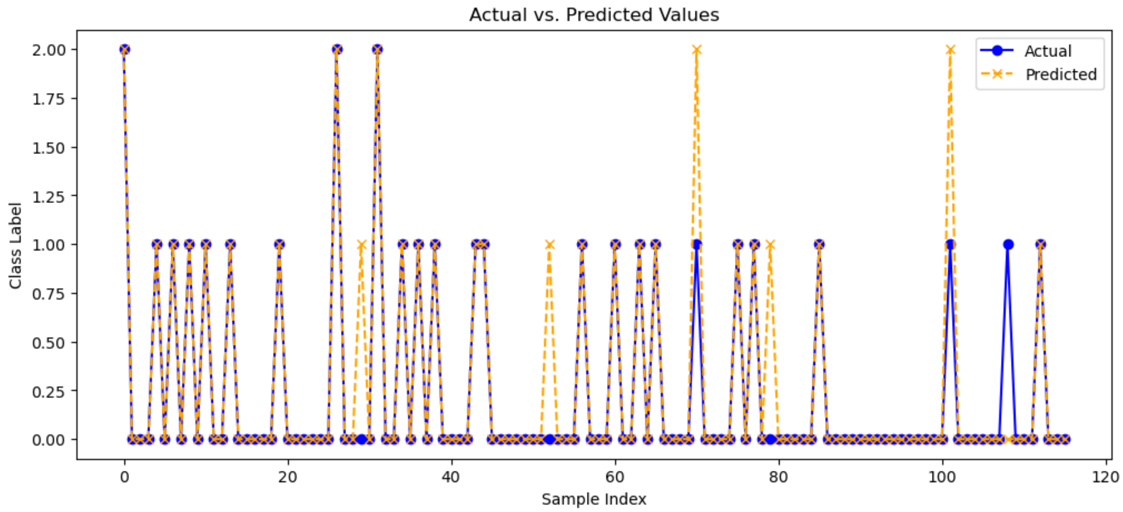

Figure 8. Prediction graph comparing the actual and predicted groundwater quality class labels generated by the Stochastic Neural Network (SNN) model for the test dataset.

Further, the model produces accurate predictions, particularly for Class 0 (excellent water), showing that it has a powerful learning source. Misclassifications occur due to overlapping attributes; yet, Class 1 (good water) is also predicted effectively. Notably, Class 2 (poor water) had perfect recall, which means that no instances of Class 2 were overlooked. The balance between sensitivity and specificity in minority class predictions is statistically supported by the classification report and visually by the prediction graph, which validates the confusion matrix findings. ROC curves of SNN for multi-class classification also demonstrate the effectiveness of the model, with high AUC scores suggesting high level of class separability. Cross-validation resulted in a consistent mean accuracy of 0.93 with moderate variability (SD = 0.03), demonstrating consistency across datasets. Also, the model's increased accuracy with rainfall inclusion shows that it successfully captured these hydrogeological relationships. Overall, the SNN model showed excellent predictive power, with encouraging prospects for environmental monitoring. By incorporating randomness during inference, the SNN model makes it possible to evaluate predictive uncertainty and more accurately reflect groundwater dynamics, an advantage that significantly aids in making sound environmental management decisions.

CONCLUSIONS

In this study, a customized Stochastic Neural Network (SNN) model was developed to categorize and predict groundwater quality in Kanyakumari District of Tamil Nadu. Using Monte Carlo Dropout and a wide range of water quality parameters, the model significantly improved generalization and quantified predictive uncertainty. Evaluation metrics such as the classification report, confusion matrix, and ROC-AUC curves, verified the model's capacity to distinguish between groundwater quality classes, which was further reinforced by consistent 5-fold cross-validation results. With 95% accuracy, the model is suitable for regions with dynamic groundwater, providing probabilistic predictions required for effective groundwater management. Future research might delve into ensemble stochastic models, Bayesian deep learning, or other physical frameworks, as well as extending the approach to real-time applications on IoT-enabled platforms. To gain a more detailed understanding of groundwater quality trends, other hydrogeological, climatological, and human-influenced variables could be included. Adding variables like seasonal fluctuations, land-use changes, and soil composition could increase the accuracy of predictions.

Acknowledgements

The authors extend their sincere gratitude to Dr. Selvam and Dr. Siva Subramanian of Department of Geology, V.O. Chidambaram College, for their valuable insights and constructive comments, which significantly enhanced the clarity and depth of this work. We also express our heartfelt appreciation to Dr. Saravanan and Dr. Arun Bose of Centre for Geotechnology, Manonmaniam Sundaranar University, for their steadfast support and encouragement throughout the study. "The authors express deep gratitude to our Vice Chancellor Prof.Dr.N.Chandrasekar for the invaluable support" Our sincere thanks are further extended to Dr. O. P. Pandey, Chief Editor of JIGU, for his thoughtful guidance and thorough review of the manuscript.

Author Credit Statement

AJ: Conceptualization, methodology, writing original draft, SY: Writing, review and supervision

Data Availability

Data cannot be made publicly available; readers should contact the corresponding author for details.

Compliance with Ethical Standards

The authors declare no conflict of interest and adhere to copyright norms.

REFERENCES

BIS, 2012. Indian standard specification for drinking water, Second Revision, IS 10500: 2012, Publication Unit, Bureau of Indian Standard, Manak Bhawan, New Delhi, pp. 1-6.

Chollet, F., 2017. Deep Learning with Python, Second Edition, Manning Publications Co., Shelter Island, New York, pp. 8.

District Statistical Handbook, 2022-2023. Department of Economics and Statistics, Kanniyakumari District, Government of Tamil Nadu, Land of Cash Crops, India. https://kanniyakumari.nic.in

Domenico, P.A. and Schwartz, F.W., 1990. Physical and Chemical Hydrogeology. Wiley, New York, pp. 824.

Emig, J., Reuter, U., Bolz, P. G., Leischner, S. and Falla, G. C., 2023. A stochastic neural network-based approach for metamodelling of mechanical asphalt concrete properties. Int. J. Pavement Engineering, 24 (1). https://doi.org/10.1080/10298436.2023.2177650

Florensa, C., Duan, Y. and Abbeel, P., 2017. Stochastic Neural Networks for Hierarchical Reinforcement Learning. arXiv (Cornell University). Proc. Int. Conf. Learning Representations (ICLR), Apr. 2017. doi: 10.48550/arxiv.1704.03012.

Gholami, V. and Sahour, H., 2022. Prediction of groundwater drawdown using artificial neural networks. Environ. Sci. Pollution Res., 29(22), 33544-33557. doi: 10.1007/s11356-021-18115-9

Giao, N. T., Nhien, H. T. H., Anh, P. K. and Thuptimdang, P., 2023. Groundwater quality assessment for drinking purposes: a case study in the Mekong Delta, Vietnam. Scientific Reports, v.13, no.1.

Gilli, E., Mangan, C. and Mudry, J., 2012. Hydrogeology: Objectives, Methods, Applications. CRC Press, pp. 4-14.

GSI, 2005. Geology and Mineral Map of Kanyakumari District (eds. Sundaram R., Ranganathan M., and Vasudevan D., compiled in OP: TNKP, Chennai; Published by Geological Survey of India, Southern Region, Hyderabad.

Haggerty, R., Sun, J., Yu, H. and Li, Y., 2023. Application of machine learning in groundwater quality modeling - A comprehensive review. Water Res., v.233, 119745

Kalariya, V., Parmar, P., Jay, P., Tanwar, S., Raboaca, M. S., Alqahtani, F., Tolba, A. and Neagu, B., 2022. Stochastic neural Networks-Based algorithmic trading for the cryptocurrency market. Mathematics, 10(9), 1456. https://doi.org/ 10.3390/math10091456

Karamouz, M., Ahmadi, A. and Akhbari, M., 2011. Groundwater hydrology: Engineering, Planning, and Management. CRC Press, p. 142

Karnena, M. K. and Saritha, V., 2019. Evaluation of Spatial Variability in Ground Water Quality using Remote Sensing. Int. J. Recent Technol. Engineering (IJRTE), 8(2), 4269-4278

Kirsch, R., 2006. Groundwater geophysics: a tool for hydrogeology, Second Edition, Choice Reviews Online, Springer, 44, pp. 511.

Kishor, K., Aggarwal, A., Srivastava, P. K., Sharma, Y. K., Lee, J. and Ghobadi, F., 2025. A Systematic Literature Review of MODFLOW Combined with Artificial Neural Networks (ANNs) for Groundwater Flow Modelling. Water, 17(16), 2375. https://doi.org/10.3390/w17162375

Kulisz, M., Kujawska, J., Przysucha, B. and Cel, W., 2021. Forecasting water quality index in groundwater using artificial neural network. Energies, *14*(18), 5875; https://doi.org/10.3390/en14185875

Lohani A.K. and Krishnan. G., 2015. Groundwater level simulation using artificial neural network in southeast, Punjab, India. J. Geol. Geophys., 4(3). DOI: 10.4172/2329-6755.1000206

Müller, A. C. and Guido, S., 2016. Introduction to Machine Learning with Python: A Guide for Data Scientists, 279-299.

Pradhan, M. and Kumar, U. D., 2019. Machine Learning with Python. Pp.149-160

Sarmadi, S., Winkle, J. J., Alnahhas, R. N., Bennett, M. R., Josić, K., Mang, A. and Azencott, R., 2022. Stochastic neural networks for automatic cell tracking in microscopy image sequences of bacterial colonies. Math. and Computational Appl., 27(2), 22. https://doi.org/10.48550/arXiv.2104.13482

Schneider, E., Dai, L., Topper, R. Q., Drechsel-Grau, C. and Tuckerman, M. E., 2017. Stochastic Neural Network Approach for learning High-Dimensional Free Energy Surfaces. Phys. Rev. Lett., 119 (15), 150601

Stylianoudaki, C., Trichakis, I. and Karatzas, G. P., 2022. Modeling groundwater nitrate contamination using artificial neural networks. Water, 14(7). 1173; https://doi.org/10.3390/w14071173

Vasanthavigar, M., Srinivasamoorthy, K., Vijayaragavan, K., Ganthi, R. R., Chidambaram, S., Anandhan, P., Manivannan, R. and Vasudevan, S., 2010. Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, Tamilnadu, India. Environ. Monitoring and Assess., 171. 595-609.

WHO, 2011. World Health Organisation. Guidelines for Drinking Water Quality, 4th ed., pp. 223-228

Received on 17-05-2025; Revised on 22-11-2025; Accepted on 29-11-2025