Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776794

SEQUENTIAL LEARNING NEURAL NETWORK BASED WATER QUALITY PREDICTION AND CLASSIFICATION

[1]S.Geetha*
Research scholar, Department of Electronics and communication,
Rathnavel Subramanian college of Arts and Science,Sulur, Coimbatore.
*Corresponding Author Mail id: smg832000@gmail.com

[2]Dr.P.Venkateswari
Associate Professor, Department of Electronics and communication,
Rathnavel Subramanian college of Arts and Science,Sulur, Coimbatore.
Mail id: venkateswari.p@rvsgroup.com

[3]Dr.T.sivakumar
Principal, Rathnavel Subramanian college of Arts and Science, Sulur, Coimbatore
Mail id: Sivakumar.t@rvsgroup.com

Abstract

Environmental degradation, particularly water contamination, has gotten significantly worse in recent years as a consequence of increased economic activity and population growth. Monitoring water quality is a critical part of water pollution prevention and management. As a fundamental human right and an integral part of sound public health policy, everyone should have access to clean drinking water. This is an concern of local importance in terms of health and development. Reductions in adverse health consequences and health care expenses outweigh the interventions' costs in some regions, indicating that water supply and sanitation expenditures might have a positive economic impact. Controlling water contamination necessitates continuous, real-time monitoring of water quality. It needs a lot of time and money to implement biological and lab-based approaches to water pollution reduction. This research introduces a new Sequential Learning Neural Network (SLNN) for the prediction and classification of water quality as a solution to this problem. There are two primary steps in the proposed model: preprocessing of data and classification of water quality. Data pre-processing takes place in a variety of ways, including data transformation, data splitting, and data normalisation, at the beginning of the process. In addition, the SLNN model makes use of several water variables to predict and classify water quality. Data includes pH, Hardness, Solids, Chloramines, Sulfate and Turbidity, is used to test the projected model's performance. The proposed model outperformed the other models in the tests, according to the findings.

Keywords:- Drinking Water; Pre-processing; Sequential Learning Neural Network; Water Pollution; Water Quality.

Introduction

Dynamic regulation of water quality and unexpected events benefit greatly from water quality prediction. It's a fundamental task in water resource management and pollution

avoidance [1]. Traditional management methods use inferior-disposal instead of prior-prevention. Several elements, some of which are not well-structured, affect water indicators [2]. In addition, the complexity of water quality prediction is determined by the nondeterministic and nonlinear feature. Predicting water quality can be done in a number of ways [3], both domestically and internationally [3]. Mathematical statistics [4], grey theory [5], chaos theory [6], and neural networks [7] make up the majority of these approaches. In modelling, the mathematical statistics method works well, although the predictions are not perfect [7]; Nonlinear functions can't be approximated using grey theory; they're too complicated for it. When the training data is very rich, the chaos theory method can be useful [8]. Traditional neural networks, with their non-linearity, concept, and learning capabilities, are ideal for dealing with nonlinear, randomised data. Time series data can't be processed by typical neural networks because of its structure [9].

Seasonal variation is evident in the time series data used to measure water quality indicators. In the field of period series forecast, water quality is included [10]. In order to deal with time series data, this research provides a water quality prediction approach that uses a neural network. It is possible to learn the features of data using deep learning, a technique of machine learning [11-12]. In recent years, researchers have been attempting to address the time series prediction problem using deep-learning-based algorithms. In the past, water quality testing was done manually by obtaining water samples and sending them to laboratories for analysis, which is a time-consuming, expensive, and time-consuming operation. It's impossible to get real-time data using these methods [13]. For continuous data transmission via wireless technology is provided by the proposed water quality monitoring system [14].

Temperature, dissolved oxygen, pH value, ammonia nitrogen, nitrites, nitrates, and other water quality factors interact with one other to create a water environment that is complex, dynamic, and non-linear [15]. Using models to predict water quality time series in recent years has made some progress. For example, most of the previous schemes used the combined model to improve the classic BP neural network and support vector machine method [16-17]. This shows that the combined model can accurately forecast water quality. As a result of this bottleneck, they are unable to learn the long-term correlation between water quality time series and the hidden information. Deep learning algorithms that can automatically extract features are becoming a research focus right now. With the help of SLNN and pre-processing models, this research attempts to significantly improve the accuracy and stability of the dissolved oxygen forecast.

The study of existing techniques with limitations is provided in Section 2. The brief explanation of proposed model with mathematical expression is given in Section 3. The validation of proposed SLNN model with existing techniques is presented in Section 4. Finally, the scientific contribution with future work is depicted in Figure 5.

2. Related Works

Hybrid decision tree-based machine learning techniques were developed by Lu, H. [18] in order to improve short-term water quality predictions. The two hybrid models' foundations are XGBoost (extensive gradient boosting) and random forest (full ensemble empirical mode decomposition with adaptive noise), both of which utilise an advanced data denoising

technique (CEEMDAN). Using 1875 data points (collected hourly) between May 1 and July 20, 2019, researchers examined the water resources of the Tualatin River Basin's Gales Creek location, which is home to one of the world's most polluted rivers. Two hybrid models are used to estimate water temperature, dissolved oxygen, pH, specific conductance, and turbidity. Dissolved organic materials is also expected to be fluorescent. Six metrics of error are used to compare the outcomes of the two models with those of the other four classic models.

Using Liu J.'s raw data, a new, massive dataset with 23,204 data points is provided [19]. After applying threshold processing, mean proximity method, wavelet filter and an improved smoothing technique to water quality indicators, data is cleaned up for further analysis and interpretation. The Pearson correlation coefficient method is used to determine the link between water quality and other dynamical variables. Data is weighted based on the correlation coefficients that have been discovered. Bi-S-SRU (Bidirectional Stacked Simple Recurrent Unit) is the name of a novel learning network that incorporates a backward SRU node into training. To build a prediction model, data must be collected. Our tests demonstrate how much more precise and fast this prediction system is than the previous one.

It was possible for Bai, W. [20] to boost the accuracy of water quality prediction by employing ensemble empirical mode decomposition (EEMD). The attention mechanism (AT) and the long-short term memory network (LSTM) were used to produce water quality prediction models based on deconstructed data (LSTM). In addition, AT can be injected before or after the LSTM network, depending on the time step dimension or feature dimensions of AT concentration. The model's predictions were tested using data from the Laokou station in the Pearl River Basin. The experiments show that the LSTM model outperforms the traditional MLP and SVR models in this regard. It is now possible to make more accurate predictions about water quality using this model now that it incorporates new features like EEMD, LSTM, and AT.

Mei, P. presents a CNN-GRU-Attention (CGA) model for predicting water quality [21]. The CGA model uses CNN, GRU, and attention layers to extract the short-term elements of water quality data while modifying the weights of individual neurons based on the correlation between neurons. CGA exceeds the individual LSTM and GRU models in terms of performance, while a lower data correlation correlates to a higher effect on the optimization of the model. This model's predictions are more accurate, but they are also more stable than those of the CNN-LSTM-Attention model. It is possible to improve the TOPSIS model by using weight correlation relations to measure water quality in total. According to the findings, summer water quality is somewhat worse than winter because of neighbouring industrial and agricultural contamination, and raw water quality is currently dropping.

Artificial neural networks (ANNs), support vector machines (SVMs), and hybrid learning algorithms were used to predict algal development and eutrophication in Hong Kong's Tolo Harbour by comparing more than 30 years of measured data, which were simulated and compared. [22] T.Deng Using these two ML techniques, it appears that algal growth trends and size may be accurately predicted. ANN outperforms SVM in terms of speed and accuracy, but at the expense of a longer training time. Furthermore, it has been shown that the ML techniques used might provide robustness to understand the complex interactions between algal dynamics and diverse coastal environmental variables and so accurately identify significant variables.

In his research, Wu, J. [23] used the grey correlation analysis method to uncover the relationships between dissolved oxygen, temperature, pH, and turbidity. It was also used to forecast dissolved oxygen levels in water. Dissolved oxygen was found to be strongly correlated with temperature and pH. The relationship between temperature and dissolved oxygen is antagonistic, while the relationship between pH and dissolved oxygen is positive. With an accuracy rate of more than 90%, dissolved oxygen, which impacts water quality, is a good predictor. The study's findings will be useful in the fight against water pollution and the management of water resources. Most of the existing works focused on collected dataset and none of the above studies used the Kaggle dataset for prediction process. To improve classification accuracy, pre-processing techniques are not focused by the existing techniques. This issues are addressed by this research model, which is explained in the next section.

3. Proposed Methodology

The basic work flow of the projected ideal is given in Figure 1.

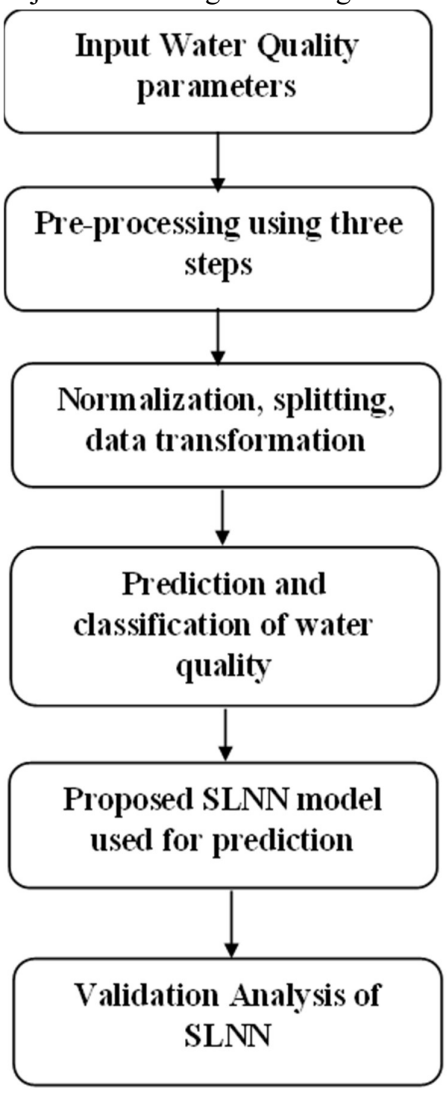

Figure 1: Workflow of projected ideal

3.1. Predicting water potability dataset

The dataset has been taken from the link : https://www.kaggle.com/adityakadiwal/water-potability

The parameters used in the dataset are:

1. Hardness
2. Solids (Total dissolved solids - TDS)
3. Organic_carbon
4. pH value
5. Chloramines
6. Sulfate
7. Conductivity
8. Trihalomethanes
9. Turbidity
10. Potability (0 = Not potable , 1 = Potable)

3.2. Pre-processing

3.2.1. Data Normalization

Transformations are needed to normalise raw data for both independent and dependent variables, as required by neural computation algorithms. To match the distribution of the estimated outcomes, the transformation adjusts input variables. The initial equalisation of the relevance of variables and the improved interpretability of network weights are the reasons for scaling data samples. Following decomposition and linear transformation, all training inputs and outputs are treated equally in neural network training and network convergence, ensuring that all training inputs and outputs receive equal attention during the training process and accelerating network convergence. Equation (1) is used to scale or normalise the data on water quality.

$$X_i = a \times \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

A and B are constants, and a and B indicate the values of the original data, respectively, in the normalised data set X_i .

Data samples are also improved by z-score transformation. In order to determine the Z-score, the following formula must be followed. This is a statistical method for determining the degree of deviation of a data value from the average. As the original data's mean and standard deviation are depicted in this equation, x represents the equation's solution.

$$X_i = \frac{x_i - \bar{x}}{\sigma} \quad (2)$$

3.2.2. Data division (data splitting)

A further step in validating water quality models requires the use of previously unobserved data in order to further bolster their predictions. To ensure that the model would perform nearly identically on a variety of datasets. This is prevalent in the literature. Predictive models' performance can be significantly impacted by how training and validation data are divided.

3.2.3. Data Transformation

To ensure interoperability with data mining methods, data transformation is utilised in the building industry to translate numerical data into categorical data. Because of their simplicity, the equal-width and equal-frequency approaches are extensively utilised. With the equal-width approach, you can break up a variable's range into numerous smaller ones. Most of the time, the user sets the number of intervals depending on his or her specific expertise in the subject matter [24]. Data can be divided into numerous equal-frequency intervals, each containing roughly the same amount of data. A variable refrigerant flow system's five numerical variables were divided into three categories using the equal frequency method. With the equal-frequency technique, outliers are less likely to be a problem.

There are many techniques for turning categorical variables into numerical ones in order to make it easier to generate forecasts for the future. Many studies have relied on a technique known as "one-hot encoding," which creates an L-1-column matrices for categorical variables with L levels. Having a lot of categorical variables might lead to a lot of data that is quite multi-dimensional. Deep learning algorithms such as embedding networks can be used to represent categorical variables using dense representations [25]. To further reduce computation costs, data transformation can be employed to analyse large-scale time series data. Simplified numerical time series in construction data can now be represented as meaningful symbol sequences thanks to the symbolic aggregate approximation (SAX) method [26, 27]. In this way, the original data could be compressed while the loss of information during data transformation is minimised.

3.3. Classification using SLNN

Minimal resource allocation network (MRAN) and GAP-RBF are two sequential learning algorithms used to predict water quality.

3.3.1. MRAN

For the input data set with L training examples, $\{(u_1, y_1), \dots, (u_t, y_t), \dots, (u_L, y_L)\}; u_t \in \Re^m, y_t \in \Re^n$the regression issue is one in which the functional relationship $f: u_t \to y_t$,is approximated to the greatest degree of accuracy achievable in order to improve the accuracy of output prediction for new samples.

The MRAN is an RBF network that uses a sequential learning technique to mimic this functional connection. With no hidden neurons at the start, MRAN grows its network by adding new neurons, pruning old ones, and changing network parameters as it gets smarter. As a result, each sample is only viewed once and then discarded. In order to maintain generality, we might assume that K more neurons are inserted once t-1 examples have been learned. Each time a sample is taken, (u_t, y_t)in the data set, the replies of the K in the hidden layer are assumed by:

$$h_k(u_t) = exp - \left(\frac{||u_t - \mu_k||^2}{\sigma_k^2}\right); k = 1, \dots K \quad (3)$$

where, $\mu_k \in \Re^m$is the k-th hidden neuron, and $\sigma_k \in \Re$is its width.

The output of the following form network:

$$\hat{y}_t == a_0 + \sum_{k=1}^{K}(a_k h_k(u_t)) \quad (4)$$

where a_k is the bias term and $\hat{y}_t$ is the anticipated output, and , a_0 is the weight connecting the kth hidden neuron to the output neuron.

New hidden neurons are assigned and the existing network parameters are adjusted during the MRAN learning process. In order to add a neuron to the hidden layer, the following conditions must be met:

$u_t - \mu_{tr} > \varepsilon_t$ (5)

$e_t = y_t - \hat{y}_t > e_{min}$ (6)

$e_{rms} = \sqrt{\frac{1}{M}\sum_{i=t-(L-1)}^{t}(y_i - \hat{y}_i)^2} > e_{,min}$ (7)

where, μ_{tr} is the hidden neuron center of the closest to u_t. The parameters ε_t, e_{min} and $e_{,min}$ are the thresholds to be designated suitably. The threshold $\varepsilon_t = \max(\in_{max} \gamma^t, \in_{min})$, where $0 < \gamma > 1$ is a deterioration constant and the length is M. As a result, the growth requirement in MRAN (5) and (6) is based on the RMS value of the output error over a sliding data window, in addition to the RAN condition (Eq. (7)). E rms over a sliding window (M) must be greater than this level for this condition to be met. Because of the growth and pruning method, the number of concealed units will change. This condition ensures that the transition is smooth. In order to add a neuron, MRAN measures the RMS error and the sample's distance from the existing neuron centres before determining whether the sample is unique. MRAN's pruning algorithm computes the normalised hidden neuron outputs for the samples. A neuron is removed from the network if its normalised output is less than a threshold value () over a series of successive samples determined by the sliding window (M)..

The MRAN can be abridged as:

1. *For each sample u_t, compute the network output $\hat{y}_t$.*
2. *If criteria (5)– (7) are satisfied, add a neuron with RBF center and width to the hidden layer.*
3. *If the contribution of a hidden neuron for a window of consecutive samples is below a threshold, delete the neuron from the network.*
4. *Adjust the centers, widths and weights of the network using an Extended Kalman Filter.*
5. *Increase t and go to step* 1.

3.3.2. Growing and pruning RBF network

Similar to the MRAN, the GAP-RBF network calculates neuronal importance using the requisite precision, but it does so in a different manner. It's possible to calculate a neuron's contribution to the network output by averaging all of its sample inputs. Additional criteria for adding neurons to a network are used by the GAP-RBF, such as the distribution of inputs in Equations 5 and 6. Adding a neuron to the hidden layer is only done if the relevance of the neuron outweighs the desired precision. A neuron must be pruned when its precision falls

below the accepted standard of significance. The Euclidean distance between the current sample and its nearest neuron is used to calculate a neuron's significance. Only if the input sample is significant enough to warrant the addition of an additional neuron are the parameters of a neuron altered.

4. Results and Discussion

One NVIDIA Tesla V100 GPU was utilised in conjunction with PyTorch software and the fastai library.

4.1. Performance metrics

Every sample is assigned a predicted label based on the classification model's predictions. As a result, each sample is classified into one of the following four categories:

- Actual positives that are properly foretold positives are named true positives (TP);
- Actual positives that are erroneously forecast negatives are named false negatives (FN);
- Actual negatives that are properly forecast negatives are named true negatives (TN);
- Actual negatives that are imperfectly forecast positives are named false positives (FP).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$MCC = \frac{TP.TN-FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \quad (9)$$

$$F_1 score = \frac{2.TP}{2.TP+FP+FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall\ /\ Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

The confusion matrix of proposed classifier is shown in Figure 2. The precision and recall curve for proposed model is given in figure 3. ROC representation of SLNN is depicted in Figure 4.

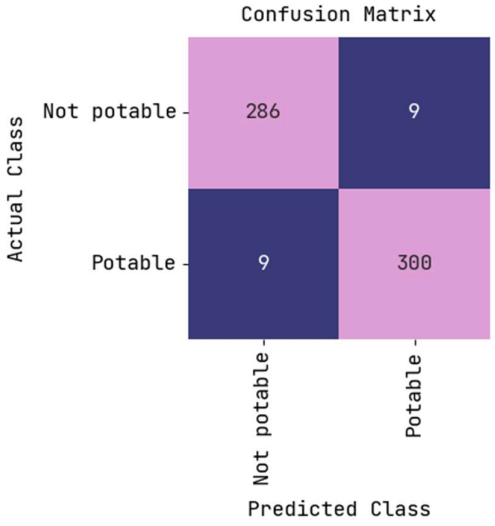

Figure 2: Confusion Matrix of proposed SLNN model

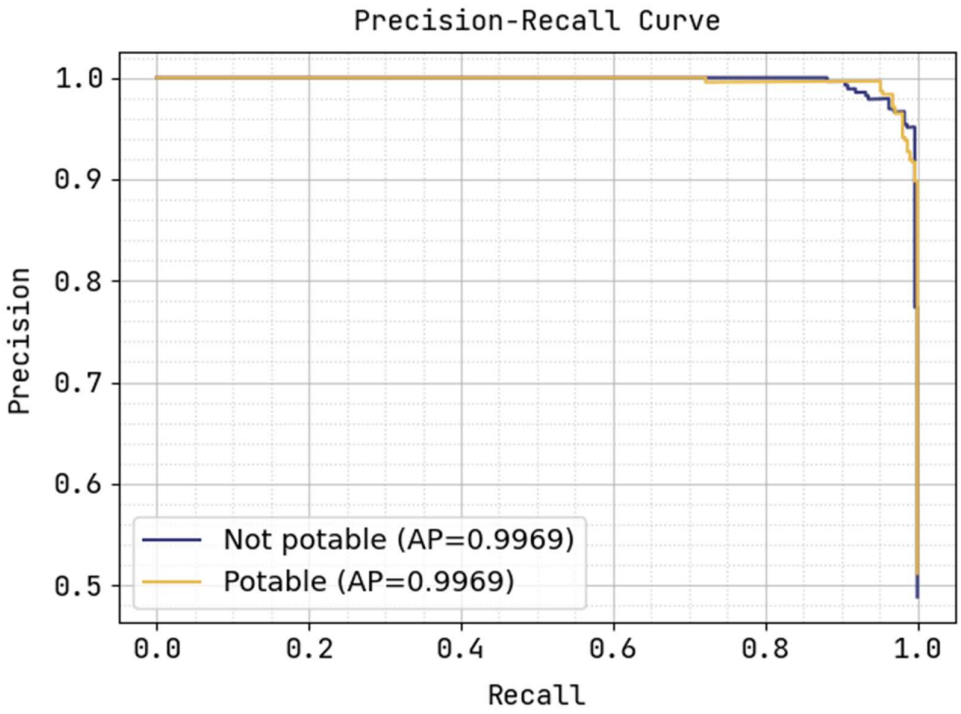

Figure 3: Precision-Recall Curve for proposed SLNN model

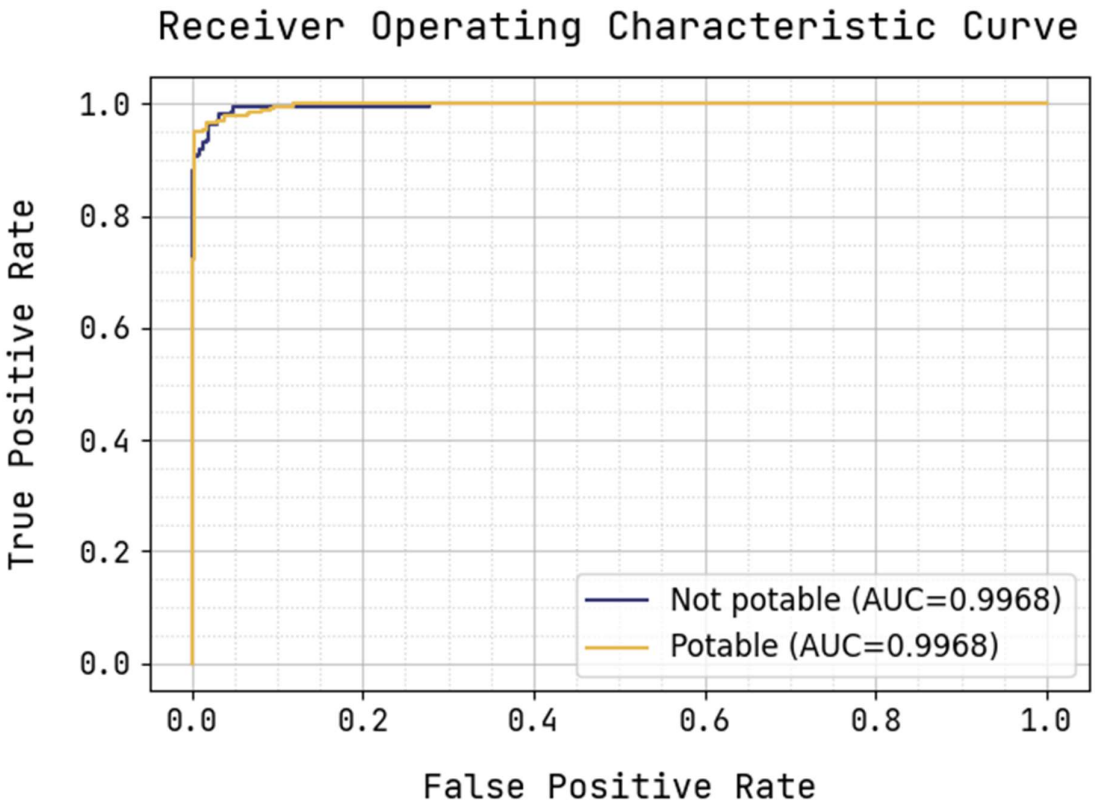

Figure 4: ROC representation of proposed SLNN model

4.3. Performance analysis of SLNN with existing techniques

The existing techniques such as XGBoost with RF [18] and ANN [22] are implemented with the input datasets in terms of various metrics, which is shown in Table 1 and Figure 5.

Table 1: Experimental Analysis of Proposed model with existing techniques

Metrics	SLNN	XGBoost-RF	ANN
Accuracy	97.02	94.56	92.89
Precision	97.02	94.65	92.89
Recall	97.02	94.65	92.89
Specificity	97.02	94.65	92.89
F1-Score	97.02	94.65	92.89
ROC-AUC Score	99.68	98.56	95.65
False Positive Rate	0.0298	0.0456	0.0856
Mathews Correlation Coefficient	94.04	90.35	86.25

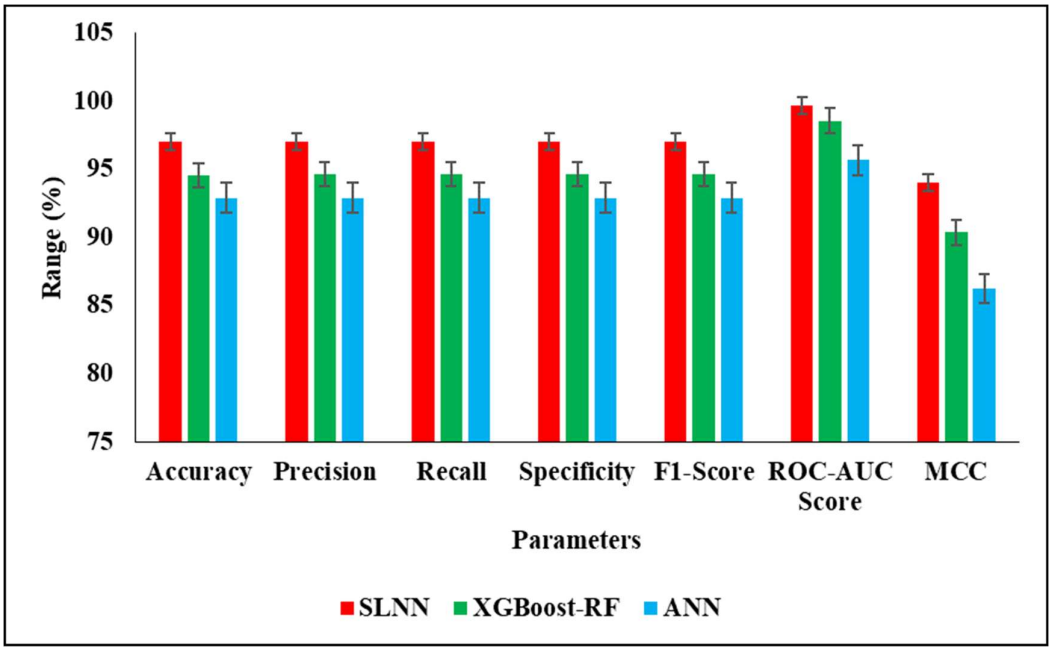

Figure 5: Graphical Representation of proposed model

In the above Table 1, the experimental analysis of proposed model is represented with existing techniques under different metrics. The three different techniques are used as SLNN, XGBoost-RF and ANN. In this analysis the Accuracy of SLNN reaches the 97.02% and the Mathews Correlation Coefficient of 94.04 % and the XGBoost-RF model reaches the accuracy value of 94.56%. Finally, the ANN model reaches the accuracy of 92.89% respectively. In the analysis of FPR, the proposed model reaches less error rate, where ANN reaches high error rate (i.e., 0.08). The ANN model achieved nearly 92% of precision, recall, specificity and F-score, XGBoost-RF achieved 94% of precision, recall, specificity and F-score. But the proposed model achieved 97% of precision, recall, specificity and F-score. In this comparison analysis, the SLNN reaches the better performance than other models.

5. Conclusion

As a way to solve this problem, a new Sequential Learning Neural Network (SLNN) is presented for predicting and classifying the quality of water. The main parts of the proposed model are data preprocessing and water quality classification. At the beginning of the process, data is pre-processed in a number of ways, such as by transforming the data, splitting it up, and making sure it is all the same size. The SLNN model also uses a number of variables about water to predict and classify water quality. The publicly available dataset is used to test the efficiency of the proposed model. The proposed model achieved 97.02% of accuracy, where the existing techniques such as ANN achieved 92.89% of accuracy and XGBoost-RF achieved 94.56% of accuracy . From this analysis, it is clearly proves that the proposed model achieved better performance. As a future work, this research model is tested with different databases by implementing the parameter optimization in the proposed model.

References

[1] Barzegar, R., Adamowski, J. and Moghaddam, A.A., 2016. Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. *Stochastic environmental research and risk assessment*, *30*(7), pp.1797-1819.

[2] Chatterjee, S., Sarkar, S., Dey, N., Sen, S., Goto, T. and Debnath, N.C., 2017, July. Water quality prediction: Multi objective genetic algorithm coupled artificial neural network based approach. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)* (pp. 963-968). IEEE.

[3] Vimal Kumar M N (2020) "A Novel Deep Learning Classifier and Genetic Algorithm based Feature Selection for Hybrid EEG-fNIRS Brain-Computer Interface", NeuroQuantology, Vol. 18, Issue 9, pp 125-134, doi: 10.14704/nq.2020.18.9.NQ20224

[4] Liu Fang and Hu Caihong, "Establishment and Application of Forecast Model for Precipitation Based on Mathematical Statistics," Meteorological and Environmental Sciences, vol. 37, 2014, pp. 89-93, doi: 10.3969/j. ssn. 1673-7148.2014.02.015.

[5] Sun Yuanhuan and Hu Yuzhuo, Application of Improved Gray Neural Network Model for Water Quality Prediction, Chongqing: Chongqing University, 2010.

[6] Sun Xihao and Yan Lei, "Prediction Model of Taihu Using Improved Fuzzy Time Series," Science and Technology&Innovation, vol. 21, 2016, pp. 15-16, doi: 10.15913/j. Cnki. kjycx.2016.21.015.

[7] Yuan Honglin and Gong Ling, "Using BP Neural Network for The Prediction of Soap River Water Quality," Journal of Security and Environment, vol. 13, Apr. 2013, pp. 106-110, doi: 10. 3969/j. issn. 1009-6094. 2013. 02.023.

[8] Sun Ruiqi, A Study on Price Forecasting of US Stock Based on LSTM Neural Network, Beijing: Capital University of Economics and Business, 2015.

[9] X. Ma and Z. Tao, "Long Short Term Memory Neural Network for Traffic Speed Prediction Using Remote Microwave Sensor Data," Transportation Research Part C Emerging Technologies. Vol. 54, May 2015, pp. 187-197.

[10] M.N.Vimal Kumar (2021) "Convolutional Neural Network Architecture Based on FPGA with Reduced Requirements for Parameters", Journal of Physics: Conference Series, Vol No. 1964, Page No. 1-6, 2021, https://doi:10.1088/1742-6596/1964/6/062082

[11] Sivaraman, E., & Manickachezian, R. (2018). Unevenness measurement using the support vector machine and dynamic multiservice load balancing with modified genetic algorithm in cloud-based multimedia system. International Journal of Computer Aided Engineering and Technology, 10(6), pp.732-747, https://doi.org/10.1504/ijcaet.2018.095210.
[12] Manoharan, J. Samuel. "Study of Variants of Extreme Learning Machine (ELM) Brands and its Performance Measure on Classification Algorithm." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 83-95
[13] Das, Brinda, Jain, P.C., 2017. Real-time water quality monitoring system using the internet of things. Int. Conf. Comput. Commun. 78–82.
[14] Sugapriyaa, Tha, Rakshaya, S., Ramyadevi, K., Ramya, M., Rashmi, P.G., 2018. Smart water quality monitoring system for real-time applications. Int. J. Pure Appl. Math. 118, 1363–1369.
[15] Uma Maheswari, V., Aluvalu, R., & Chennam, K. K. (2021). Application of machine learning algorithms for facial expression analysis. Machine Learning for Sustainable Development, 9, 77.
[16] Samuel Manoharan, J. (2019), "An improved safety algorithm for artificial intelligence enabled processors in self driving cars", Journal of artificial intelligence, 1(2): 95 – 104
[17] Bedri, Z., Corkery, A., O'Sullivan, J.J., Deering, L.A., Demeter, K., Meijer, W.G., O'Hare, G. and Masterson, B., 2016. Evaluating a microbial water quality prediction model for beach management under the revised EU Bathing Water Directive. Journal of Environmental Management, 167, pp.49-58.
[18] Lu, H. and Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, p.126169.
[19] Liu, J., Yu, C., Hu, Z., Zhao, Y., Bai, Y., Xie, M. and Luo, J., 2020. Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network. IEEE Access, 8, pp.24784-24798.
[20] Bai, W., Bai, Y. and Li, Q., 2021. Study on EEMD-LSTM Water Quality Prediction Model with Attentional Mechanism. Frontiers in Science and Engineering, 1(7), pp.33-41.
[21] Mei, P., Li, M., Zhang, Q. and Li, G., 2022. Prediction model of drinking water source quality with potential industrial-agricultural pollution based on CNN-GRU-Attention. Journal of Hydrology, p.127934.
[22] Deng, T., Chau, K.W. and Duan, H.F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. Journal of Environmental Management, 284, p.112051.
[23] Wu, J. and Wang, Z., 2022. A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. Water, 14(4), p.610.
[24] Li, G., Hu, Y., Chen, H., Li, H., Hu, M., Guo, Y., et al. (2017). Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. Appl. Energy 185 (1), 846–861. doi:10.1016/j.apenergy.2016.10.091
[25] Fan, C., Sun, Y., Zhao, Y., Song, M., and Wang, J. (2019a). Deep learning-based feature engineering methods for improved building energy prediction. Appl. Energy 240, 35–45. doi:10.1016/j.apenergy.2019.02.052

[26] Piscitelli, M. S., Brandi, S., Capozzoli, A., and Xiao, F. (2021). A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. Build. Simul. 14, 131–147. doi:10.1007/s12273-020-0650-1

[27] Piscitelli, M. S., Mazzarelli, D. M., and Capozzoli, A. (2020). Enhancing operational performance of ahus through an advanced fault detection and diagnosis process based on temporal association and decision rules. Energy Build. 226, 110369. doi:10.1016/j.enbuild.2020.110369