

Improving Water Quality Assessment through Anomaly Detection
Using Hybrid Convolutional Neural Network Approach

S.V.Jansi Rani*, A. M. Ramakrishnan, K. Rishivardhan

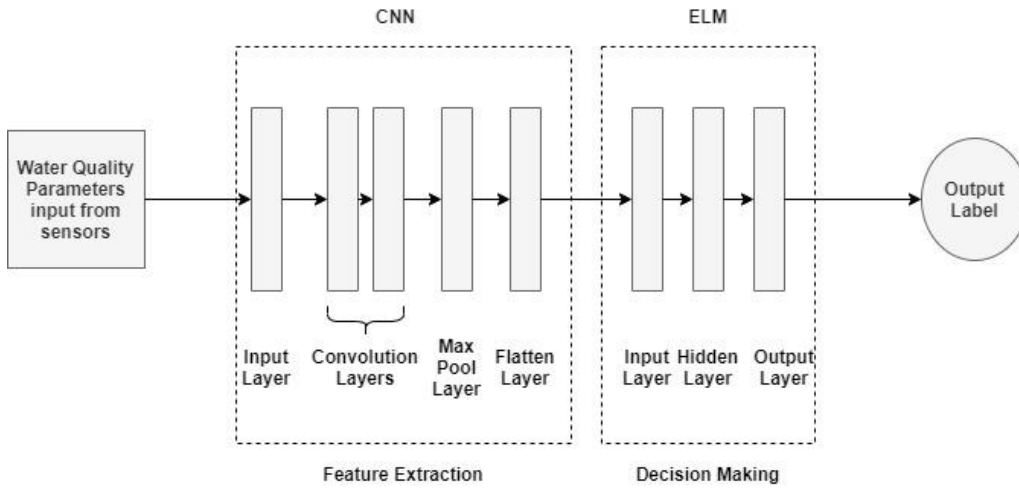
Dept of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

*Corresponding author:

E-mail: svjansi@ssn.edu.in

ACCEPTED MANUSCRIPT

GRAPHICAL ABSTRACT



ABSTRACT

Water being a precious commodity for every person around the world needs to be quality monitored continuously for ensuring safety whilst usage. The water data collected from sensors in water plants are used for water quality assessment. The anomaly present in the water data seriously affects the performance of water quality assessment. Hence it needs to be addressed. In this regard, water data collected from sensors have been subjected to various anomaly detection approaches guided by Machine Learning (ML) and Deep Learning framework. Standard machine learning algorithms have been used extensively in water quality analysis and these algorithms in general converge quickly. Considering the fact that manual feature selection has to be done for ML algorithms, Deep Learning (DL) algorithm is proposed which involve implicit feature learning. A hybrid model is formulated that takes advantage of both and it is data invariant too. This novel Hybrid Convolutional Neural Network (CNN) and Extreme Learning Machine (ELM) approach is used to detect presence of anomalies in sensor collected water data. The experiment of the proposed CNN-ELM model is carried out using the publicly available dataset GECCO 2019. The findings proved that the model has improved the water quality assessment of the sensor water data collected by detecting the anomalies efficiently and achieves F1 score of 0.92. This model can be implemented in water quality assessment.

Keywords: Anomaly Detection, Convolutional Neural Network, Extreme Learning Machine, Machine Learning, Water Quality

1. Introduction

Water is always projected to be a vital need in our everyday life. Nearly \$184 billion is spent towards supply of clean water globally (Sensus, 2013). There are serious diseases caused due to polluted water and can be classified as water-related, waterborne, water-based and water scarce diseases (NHMRC, 2011; Waterwise Rand Water, 2017). Hence, it is very important to find whether the water is of good quality. In addition to drinking water, other industries also utilize water such as gas distribution plants, oil refineries and smart grids. All water treatment and water distribution plants are generally termed as Critical Infrastructure (CI). These CIs are subjected to cyber-attacks of injecting false data (Exida, 2015) into the system. Such anomalies in data need to be found out. It becomes an inherent duty to assert the safety and quality of water being used in various use cases including drinking, sanitation and other related uses. Improving water quality has a direct and significant impact on public health (Yang et al., 2014). Therefore considering the abundant requirement of water in a global platform it is essential to detect any significant change in water conditions beforehand to avoid any burden on health care facilities. Water conditions are routinely monitored by sensors (Zhang et al., 2016) that record operative and quality indicators such as temperature, pH, etc. Any significant change in the sensor readings implies an anomaly in water quality. These anomalies serve as important indicators in a Water Distribution System (WDS) for early detection of undesirable changes and to prevent any from happening.

Anomaly Detection refers to the process of finding undiscovered patterns that differ significantly from the normal or most likely behavior in a dataset. Abnormal conditions like improper quantity of chlorine in water sample, water with high acid or base content etc., represent anomaly. In addition to these abnormal events, faulty components in water quality assessment environment also represents anomaly. Also, the water quality data might have gone through intentional or unintentional attacks. Anomaly detection can be thought of as a classification problem wherein we determine whether a particular water sample as 'normal' or 'anomalous'. There have

been multiple approaches involving Machine Learning and Deep Learning frameworks to detect anomalies in water quality data.

Water quality data is used in Water Distribution System(WDS), water quality research units etc. Though anomaly is not a frequently occurring phenomenon, if not treated properly many serious problems will be unnoticed which affects the water quality assessment. The serious problems are contamination of water and faulty operation of water quality assessment equipment etc. Also, since anomalies are not normal events which occur frequently, water quality datasets typically contain a very low proportion of anomalous water quality observations. With a heavy imbalance in dataset, it becomes tougher to train machine learning or deep learning models to detect anomalies in water quality datasets. Those models may tend to predict even anomalous data samples as normal more often, unless the models are fine-tuned to counter the imbalance. Hence it can be understood that the anomaly present in the water data seriously affects the performance of water quality assessment.

1.1. Related Work

Initially anomaly detection in Water Distribution Systems (WDS) (Shalyga et al., 2018; Muharemi et al., 2019) involved traditional approaches like sampling the water used for drinking and performing various complex analysis techniques on the collected samples in laboratories. The mentioned traditional approach using Water Quality Index (WQI)(Arivarasi et al 2017) is not only time consuming, but also involves manual labor for physical observation and analysis. Due to the complicated techniques involved for estimation, it is not appropriate to be translated to real-time detection.

The advancement of Wireless Sensor Networks (WSN) paved the way for the use of wireless sensors to record various parameters (like pH, turbidity, conductance, etc.) of water supplied to the general public (Akyildiz et al., 2002). This approach not only turned out to be time-efficient, but was also cost-effective. Monitoring using WSN involves placement of sensors at monitoring stations of WDS to capture continuous data.

With the humongous data collected by WSNs, it is vital to make sense out of the data which is to understand the data pattern and apply the inferred knowledge to detect anomalies in water. Traditionally, various simple threshold based detection algorithms were employed. Those methods did not provide satisfactory results. Anomaly detection approach was employed to identify any unusual patterns in water meter readings, not just threshold markers. Various Machine Learning (ML) algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Logistic Regression (LR)(Muharemi et al., 2018) were proposed to be employed in anomaly detection. Both SVM and LR are basically binary classifiers which are able to classify water as anomalous or normal.

Real test bed data generated from Festo MPA (Robles-Durazno et al., 2021) workstation rig and some features which are engineered are current, voltage and power from the sensors. Multilayer perceptron, k-Nearest Neighbor (KNN), SVM, Decision Trees (DT), and Random Forest (RF) are used to classify the anomalous data produced by cyber-attacks. The ML models created using ML techniques were able to achieve feasible and satisfactory results. Additionally, the efficacy of Ensemble methods (Chen et al., 2018) was experimented and they produced satisfactory results. A dynamic ensemble selection method called non-dominated local class specific accuracy was developed (Ribeiro et al., 2020). The authors had given a solution to solve problems related to false positives and false negative in water anomaly. This method benefits both consumer and water companies. A comprehensive study on the various methods is also available (Dogo et al, 2019).

The tendencies of water quality variation is found by an autoregressive method using Bayesian autoregressive (BAR)(Liu et al., 2020) and Isolation Forest (IF) algorithm is used to detect water anomaly. Both algorithms are combined to find anomalies in Potomac River of West Virginia, USA. This also provides early warning of emergency operations in advance.

A Multilayer perceptron model was used for finding anomaly in Secured Water Treatment (SWaT) dataset (MR et al., 2020). They used unsupervised approach to protect any critical infrastructure from cyber-attacks. The abnormal deviation in sensor values which is used to detect

anomalies is found by the Cumulative SUM (CUSUM) approach. The analysis is done for normal operation and direct and stealthy attacks. A framework designed to find anomalies in the high frequency data collected from situ sensors (Leigh et al., 2019) situated in rivers flowing into Great Barrier Reef. There are some prioritized anomalies such as level shifts, isolated spikes in the water data. They can be identified by the regression methods, but it suffers from false positive rates. So, feature based methods are used which gave better results but again, it suffers from false negative rates.

Rule based methods are used to figure out missing values. A combined approach of these three methods is used to form an optimal framework in detecting anomalies which produced obviously optimal outcomes. Though the performance of ML approaches is considerable, they require the need of extensive feature engineering for comparable results. To solve the computational complexity posed by explicit feature engineering, application of Deep Learning (DL) techniques (Muharemi et al., 2019; Fehst et al., 2018) were exploited in anomaly detection that performed feature extraction implicitly. Generative adversarial networks (Wu et al., 2021) can be used for detecting anomalies for underwater gliders. The model was trained on a time series two healthy dataset and then tested on nine deployment datasets. The obtained model was robust in nature. Also, various other DL algorithms like Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN), were applied in water quality detection. However unlike ML algorithms like SVM and LR, deep learning methodologies have produced unsatisfactory results. The performance degradation of deep learning models is attributed to the imbalance problem associated with water quality anomaly data sets. Subsequently Qian et al. considered the class imbalance problem affecting deep learning methods (Qian et al., 2020) and performed sampling of data before training a LSTM network. However the authors performed evaluation on a different data set and hence their work is hard to be generalized.

While existing approaches typically involve standalone Machine Learning models or standalone deep learning models, in this paper, we propose a hybrid model of combining CNN and

ELM. The innovation applied here is that we combine and exploit the advantages of both the models - quick learning time of ELM and the implicit feature extraction ability of CNN for finding anomalies in the water data. With computations on publicly available GECCO water quality dataset (SPOTSeven Lab, 2019), it is shown that the complementary strengths of ELM and CNN together is beneficial in a real-time environment where the inference time is pivotal.

The anomaly detection of the hybrid model is used for improving the water quality assessment. Some of the advantages include implicit feature learning (making it dataset invariant), less inference time and higher chances of detecting anomalies which is evident from the evaluation scores as mentioned in section 3. Since the chances of detecting anomalies are increased, it makes it possible to detect damage of critical structure of water distribution and assessment plant thereby reducing the chances of incurring huge loss. Also, the model makes it possible to detect contaminated water thereby reducing the risk of general public getting access to unsafe water.

2. Materials and methods

2.1. Dataset

The time-series collection of water features were obtained from the publicly available GECCO 2019 Industrial Challenge dataset (SPOT Seven Lab 2019). The dataset comprises of 8 indicators as mentioned in the table 1, obtained by sensor readings in a WDS from July 1, 2017 to September30, 2017. The dataset has a total number of 132480 records out of which 132268 records are non-anomalous and 212 records are anomalous. The dataset has multiple records with the values of one to many fields not filled with data. The missing values in the dataset are replaced using Multiple Imputation by Chained Equations (MICE) (Azur et al., 2011).

Also, after comparing the number of data samples for the two cases, it is observed that a significant difference in the proportion of records. The dataset is thereby concluded to exhibit a data imbalance problem. To study the effect of this high imbalance, the performance comparison of the hybrid CNN-ELM model by training it with three differently sampled datasets - the original dataset,

the oversampled dataset and a combined oversampled and under sampled dataset. For the oversampled dataset, imputing values to the minority class to the original dataset by implementing SMOTE (Bowyer et al., 2011) sampling technique. For the combined oversampled and under sampled dataset, once the dataset is SMOTE oversampled, Random under sampler is used to delete some records of the majority class according to the sampling strategy provided.

Table 1. Indicators in the dataset used

Indicator	Description
Time	Time at which the measurement was made (in yyyy:mm:dd HH:MM:SS format)
Tp	Temperature of water (in °C)
pH	pH value of water
Cond	Electrical Conductivity of water (in S/m)
Turb	Turbidity of water (in FTU)
SAC	Spectral Absorption Coefficient (in 1/m)
PFM	Pulse Frequency Modulation (in Hz)
EVENT	Boolean marker which indicates whether sample is anomalous

2.2 Algorithms:

2.2.1 Convolutional Neural Network (CNN)

CNN is very popular and makes it different among others in terms of its greater performance with audio signal, speech signal and images. The convolutional, pooling and fully connected layer are the backbone of CNN. The first convolutional layer being representative of the core operation of the architecture performs dot product of two vectors, where one is the kernel and the other in the input. The kernel slides across the input vector in one dimension (since the input is sensory data).

The sliding size of the kernel is called as stride. The filters determine the dimensionality of the output space. The kernel size will be lesser than that of input vector, so that we store only the significant and meaningful information.

Max pooling layer down-samples the input fed by taking the maximum value of the spatial window of the specified pool-size. The window gets shifted according to the specified strides. The flatten layer flattens the input fed (convert input into 1-D array) without affecting the batch size. CNN makes use of a mathematical operation called convolution to derive a weighting function $w(m)$ from every CNN layer. In $w(m)$, 'a' represents the age of measurement of time series data. CNN architecture is shown in fig 1.

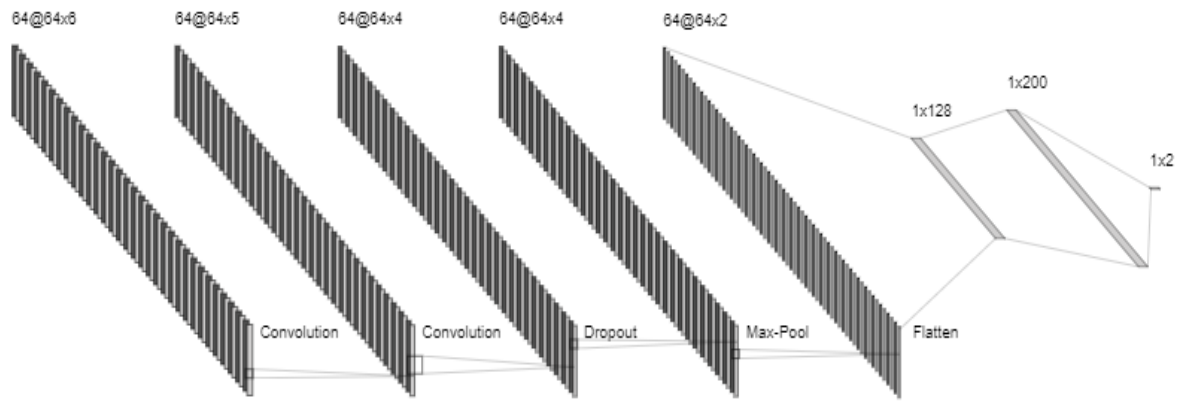


Figure 1 Architecture of CNN

The convolution operation is represented as,

$$s(t) = (x * w)(t) = \sum_{m=-\infty}^{\infty} x(m)w(t - m)$$

Here, kernel or filter is represented as w , input as x and the output as s for continuous time series data t . The experiment on CNN is done for considering its effect to extract meaningful features from data such that the computation becomes efficient.

2.2.2 Extreme Learning Machine (ELM):

Figure 2 shows the ELM architecture which comprises of three-layer feed forward architecture. The first, second and third layer corresponds to input, hidden and output layers. The weight connections

between the first layer and the hidden layer are randomly generated, set and fixed. Between the hidden and the input layer, the connection weights are trained in a single pass of input using Moore Penrose pseudo-inverse method.

The outputs from the hidden layer are produced and output is obtained from the final output layer by using the appropriate activation functions. Attributing to the good generalization performance and faster training time by ELM, the experiment is carried out on different activation functions for our task.

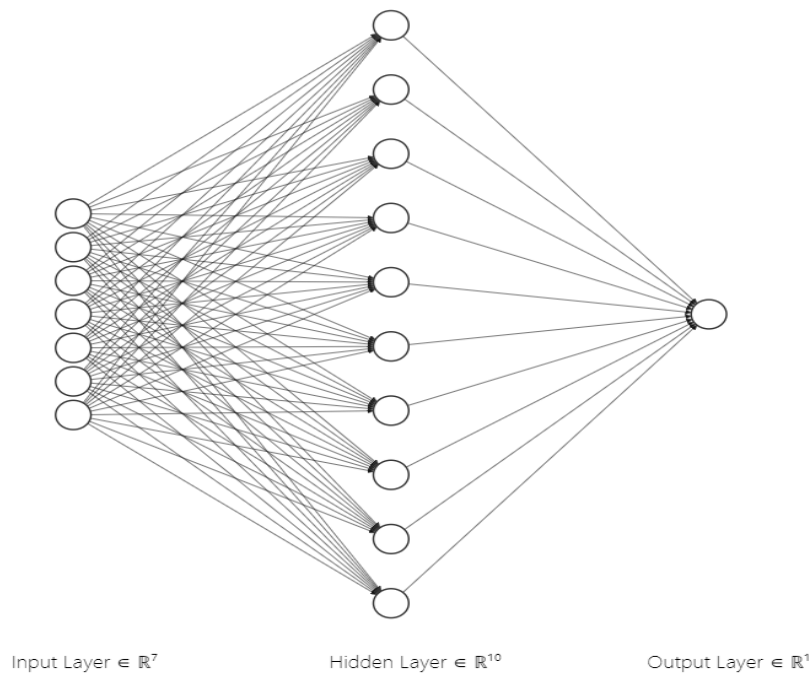


Figure 2 Architecture of ELM

2.3 Proposed Architecture

The proposed hybrid model targets to utilize the best characteristics of both CNN and ELM in such a way that CNN performs implicit feature extraction and ELM pools the trained features by CNN to effectively detect or classify records to produce the desired output as shown in fig 3. In detail, the CNN model is trained end-to-end to classify the records as anomalous or not. After

training the CNN model for a desired number of iterations, the output from the penultimate layer of the model, i.e., the flatten layer is used as input for the ELM input layer. The features extracted by CNN model is fed as input to the ELM input layer obtains from the flatten layer and uses it to train its hidden layer weights. The output finally obtained from the ELM is the target variable that denotes whether a record is anomalous or not.

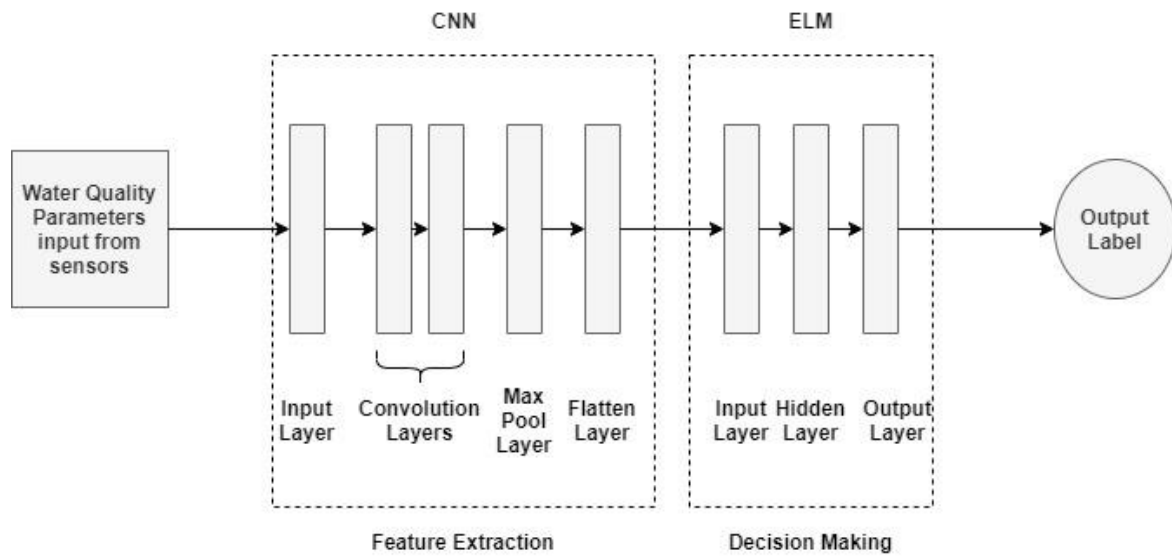


Figure 3 Hybrid CNN-ELM Architecture

The newness of this approach is that this method removes the need for manual feature extraction which is an essential step in training machine learning models. CNN does the implicit feature extraction while ELM does fast learning to predict results thereby reducing inference time. The scientific progress achieved through this approach is that this model is invariant to datasets i.e., since a deep learning model is employed to do implicit feature extraction, our model is not dataset specific and will perform well for different datasets which may contain different features. For example, GECCO 2018 and GECCO 2019 datasets had different sets of features but our model is suited to perform for both datasets. This ability of the model makes it a perfect fit for invariant datasets to produce results with high reliability in a short span of time.

3. Results and Discussion

3.1. Performance Metrics

The performance metrics such as accuracy, recall, precision and f1 score are used to evaluate the proposed hybrid model. When performing the classification task on the test set of the model, the following four outcomes are possible. Here positive means anomaly data and negative means non-anomaly data.

True Positive (TP): When both actual and predicted values are anomaly.

False Positive (FP): When the predicted value is anomaly and actual value is non-anomaly.

True Negative (TN): When both actual and predicted values are non-anomaly.

False Negative (FN): When the predicted value is non-anomaly and actual value is anomaly.

Confusion Matrix is an $n \times n$ matrix where n denotes the number of classes. The confusion matrix is a 2×2 matrix filled with actual class values against predicted class values since there are two classes. Depending upon the number of TP, FP, TN and FN the evaluation of the performance of the model is done using the following metrics:

3.1.1 Accuracy

Accuracy is the ratio of the number of correct predicted data to the total number of predicted data. Since, an imbalanced dataset is under consideration with almost 99.57 percent samples of the dataset belong to class 0, the accuracy is close to 1.0. Therefore, accuracy does not capture all the aspects of model's performance and hence do not use accuracy as a performance metric.

$$Accuracy = \frac{TN + TP}{TP + FN + TN + FP}$$

3.1.2 Precision

Precision is the ratio of number of true positives samples to the number of samples predicted as positive by the model. Precision helps us to understand how the model is able to predict and produce valid and relevant results. The formula to find precision is given below:

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

3.1.3 Recall

Recall is the ratio of number of true positives to the number of samples that are actually positive. Recall is also known as True Positive Rate (TPR). Recall helps us to understand how complete our model is predicting the true positives. The formula to find recall is given below:

$$Recall(RC) = \frac{TP}{TP + FN}$$

3.1.4 F1-score

F1-score is the harmonic mean of precision and recall. F1-score is the most important performance metric as it captures the performance of the model by taking both false positives and false negatives into account. Hence, F1-score is the most significant performance metric when dealing with a highly imbalanced dataset.

The formula to find the F1-score of the model is given below:

$$F1\ score = 2 \times \left(\frac{PR \times RC}{PR + RC} \right)$$

Since f1-score is a measure of both precision and recall, it is suitable for deciding on the performance of the model.

3.1.5 AUC-ROC

ROC (Receiver Operating Characteristic) is a probability curve plotted with the True Positive Rate (TPR) against the False Positive Rate (FPR). 'Area under the Curve' (AUC) of the ROC represents the degree of separability of the different classes by the model. Higher the AUC-ROC, the model is better at classifying a sample of class 0 as class 0 and a sample of class 1 as class 1.

$$FPR = \frac{FP}{TN + FP}$$

It is already mentioned that recall is also known as TPR. So once TPR vs. FPR is plotted, the value of AUC-ROC can be found out. For an ideal model, the AUC-ROC is 1.0. There is a need to maximize AUC-ROC so that the model performs better in distinguishing the two classes.

3.2 Implementation Results

3.2.1 Convolutional Neural Network

The standalone CNN architecture was experimented for different loss functions and optimizers. From extensive experimentation, using Binary Cross Entropy as loss function and Stochastic Gradient Descent (SGD) as optimizer provided the best performance. The experimental results for the differently sampled datasets with Binary Cross Entropy Loss function and SGD optimizer are detailed in Table 2 given below.

Table 2: Performance of Standalone CNN

Dataset	Precision	Recall	F1-score	AUC-ROC
Original	1.0	0.65	0.73	0.64
SMOTE Oversampled	0.51	0.94	0.49	0.92
SMOTE Oversampled and Random Under sampled	0.60	0.97	0.67	0.93

From table 2, it is observed that the model performs the best for the original dataset with the F1-score of 0.73.

3.2.2 Extreme Learning Machine

The standalone ELM architecture was experimented on different activation functions, i.e. Sigmoid, Tanh and Radial Basis Function (RBF) for different neuron capacity. Table 3 presents experimental results of performance of ELM based on different approaches. ELM based on 1000 neurons combined with sigmoid activation performs comparatively better than other approaches. It is observed that the model performs the best for the original dataset with the F1-score of 0.88.

Table 3: Performance of Standalone ELM

Dataset	Precision	Recall	F1-score	AUC-ROC
---------	-----------	--------	----------	---------

Original	0.98	0.82	0.88	0.82
SMOTE Oversampled	0.60	0.95	0.66	0.95
SMOTE Oversampled and Random Under sampled	0.72	0.94	0.80	0.93

3.2.3 Hybrid CNN-ELM

The combination of the best performing CNN and ELM model as observed from Table 2 and 3 is used to formulate the Hybrid CNN-ELM model. The ROC curves for the Hybrid CNN-ELM models trained with original dataset, SMOTE (Chawla et al 2002) dataset and the SMOTE oversampled and random under sampled datasets are shown in the figures 4, 5 and 6 respectively.

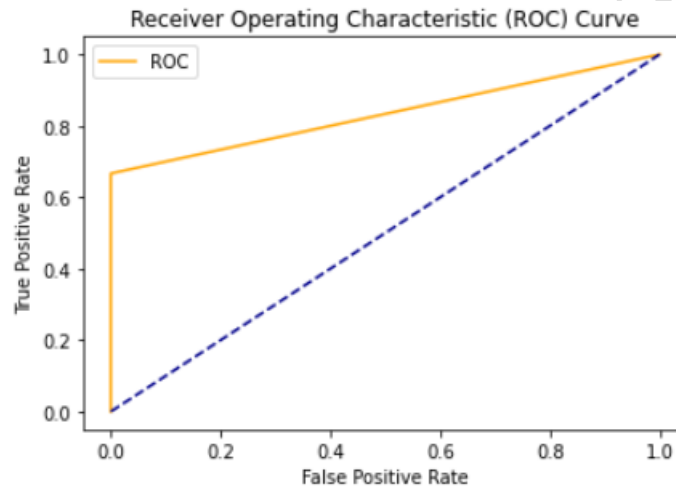


Figure 4 ROC Curve of Hybrid CNN-ELM trained with original dataset (AUC = 0.84)

It is observed that SMOTE oversampled dataset model performs significantly better with a near perfect ROC curve with AUC of 0.97 because the model is trained with relatively more samples when compared with the model trained with the original dataset.

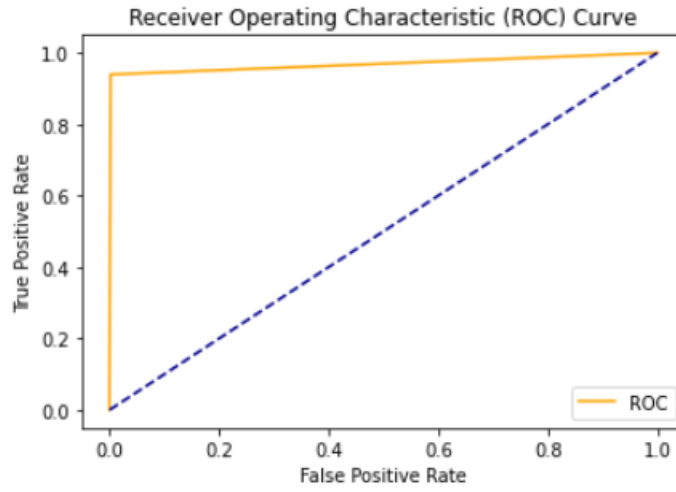


Figure 5 ROC Curve of Hybrid CNN-ELM trained with SMOTE oversampled dataset (AUC = 0.97)

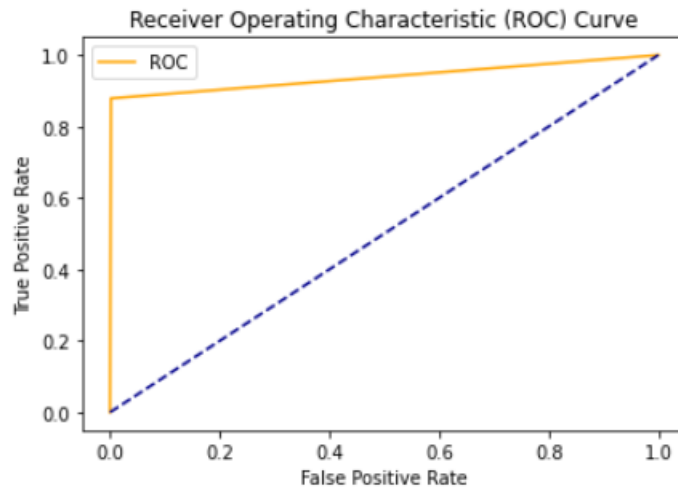


Figure 6 ROC Curve of Hybrid CNN-ELM trained with SMOTE oversampled and random under sampled dataset (AUC = 0.94)

Table 4: Confusion matrix for Hybrid CNN-ELM trained with original dataset

	Predicted Positive	Predicted Negative
Actual Positive	23 (TP)	10 (FN)
Actual Negative	0 (FP)	26463 (TN)

Table 5: Confusion matrix for Hybrid CNN-ELM trained with SMOTE oversampled dataset

	Predicted Positive	Predicted Negative
Actual Positive	31 (TP)	2 (FN)
Actual Negative	49 (FP)	26414 (TN)

Tables 4,5 and 6 show that the CNN-ELM model trained with original dataset has FP+FN=10 and has very little chances of misclassifying. In table 7, the results of proposed work of hybrid model under different datasets used for training. Considering F-1 score is a pivotal metric in the case of imbalanced datasets, the Hybrid CNN-ELM model trained with the original dataset outperforms the other two sampling based approaches.

Table 6: Confusion matrix for Hybrid CNN-ELM trained with SMOTE oversampled and random under sampled dataset

	Predicted Positive	Predicted Negative
Actual Positive	29 (TP)	4 (FN)
Actual Negative	39 (FP)	26424 (TN)

Table 7: Performance of Hybrid CNN-ELM

Dataset	Precision	Recall	F1-score	AUC-ROC
Original	1.0	0.85	0.92	0.84
SMOTE Oversampled	0.69	0.97	0.77	0.97
SMOTE Oversampled and Random Under sampled	0.71	0.94	0.79	0.94

F1-score is our major performance metric. Unlike other set of classification problems where accuracy is of more importance we focus on F1-score since our dataset is highly imbalanced and our focus is more on finding anomalies. In this case we want to maximize the true positives (anomalies) and minimize false positives and false negatives. Precision and Recall are two metrics which are usually complementary. Precision focuses on how our model is able to produce valid

results while recall focuses on how complete our model is. Though they are complementary, when true positives are higher, both precision and recall will improve. F1-score being the harmonic mean of precision and recall thereby becomes the most significant metric through which the model's performance can be determined. As shown in table 7, our model can achieve an F1-score of 0.92 which corresponds to the fact that the hybrid CNN-ELM model is able to detect anomalies more often. This proves that the model can detect anomalies caused by any unusual events injected due to intentional or unintentional attacks or failure of component. This ability of the model can be credited to the implicit feature learning ability of CNN and generalizability of ELM.

3.3 Comparison with existing approaches

Table 8: Comparison of existing methods on GECCO Water Quality Dataset

Approaches	Method	F1-score
Muharemi et al. 2018	LR	0.584
	SVM	0.087
	ANN	0.019
Muharemi et al. 2019	RNN	0.834
	LSTM	0.902
Chen et al. 2018	LR	0.480
	LSTM	0.800
Proposed Hybrid ELM Model	Hybrid CNN – ELM	0.920

Table 8 details the comparison of experimental results obtained by related works using deep learning for automatic feature selection and machine learning methods for quick convergence. In the case of machine learning based approaches, SVM benchmarks a F1 score of 0.989. In the case of deep learning approaches, our work Hybrid CNN-ELM outperforms most of the existing deep

learning works and achieves a F1 score of 0.92. There is a performance improvement of 10%, 2% when compared to Muharemi et al. 2018's RNN and LSTM approaches.

3.4 Comparison with different datasets

To prove that our model works across different datasets and it has an inherent data invariant behavior, we have used our model for another dataset GECCO 2018 (SPOT Seven Lab 2018). The results are shown in table 9. For both datasets, our model is able to achieve F1-score more than 0.9. For GECCO 2018 our model has given F1-score as 0.98 and AUC-ROC as 0.97. There is a significant increase in both the scores because of the fact that GECCO 2018 has different features and the proportion of anomalous data is more than GECCO 2019.

Table 9: Performance of Hybrid CNN-ELM when trained with different datasets

Dataset	Precision	Recall	F1-score	AUC-ROC
GECCO 2019	1.0	0.85	0.92	0.84
GECCO 2018	1.0	0.97	0.98	0.97

The significant scientific contribution of our approach is that it is data invariant and it has implicit feature learning capability which is evident from the results presented in Table 8,9.

4. Conclusion

In this paper, an improvement to water quality assessment is done through detecting anomalies in water quality data. A Hybrid CNN-ELM model is developed for improving the water quality assessment. This hybrid framework combines the implicit feature extraction ability of CNN and generalizability of ELM to make it efficient than the machine learning models which makes manual feature selection. The experiments and the results obtained is evident that hybrid framework outperforms the existing methods with f1 score of 0.92. The proposed hybrid model also works well across another dataset GECCO 2018 with f1 score of 0.98 which proves that it is data invariant.

One of the limitations is that it requires a longer time to train the model with the training time proportional to the size of the water quality dataset. Another limitation is that the model is a heavyweight component which makes it harder to import on the end hardware device. Considering the advent of optimization of deep learning models making them capable of running on an edge device (Chen, Ran, 2019), future scope of our work can be extended to deploying our model to water quality detection environment and validate real-time performance.

References

- Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer networks*, 38(4), 393-422.
- Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, 20(1), 40-49.
- Arivarasi, R. and Ganesan, M. (2017). Seasonal variation in ground water quality and its suitability for drinking and agriculture-a case study in Kancheepuram region, Tamil Nadu, India. *Global Nest Journal*, 19(1), 131-139.
- Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- Chen, J. and Ran, X. (2019). Deep Learning With Edge Computing: A Review. *Proc. IEEE*, 107(8), 1655-1674.
- Chen, X., Feng, F., Wu, J. and Liu, W. (2018, July). Anomaly detection for drinking water quality via deep biLSTM ensemble. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 3-4).
- Dogo, E. M., Nwulu, N. I., Twala, B. and Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water Journal*, 16(3), 235-248.

- Exida, L. L. C. (2015). Repository of Industrial Security Incidents (RISI) Online Incident Database. <http://www.risidata.com/Database> (Date of access: Aug 5,2021).
- Fehst, V., La, H. C., Nghiem, T. D., Mayer, B. E., Englert, P. and Fiebig, K. H. (2018, July). Automatic vs. manual feature engineering for anomaly detection of drinking-water quality. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 5-6).
- Leigh, C., Alsibai, O., Hyndman, R. J., Kandanaarachchi, S., King, O. C., McGree, J. M. and Peterson, E. E. (2019). A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of The Total Environment*, 664, 885-898.
- Liu, J., Wang, P., Jiang, D., Nan, J. and Zhu, W. (2020). An integrated data-driven framework for surface water quality anomaly detection and early warning. *Journal of Cleaner Production*, 251, 119145.
- MR, G. R., Somu, N. and Mathur, A. P. (2020). A multilayer perceptron model for anomaly detection in water treatment plants. *International Journal of Critical Infrastructure Protection*, 31, 100393.
- Muharemi, F., Logofătu, D. and Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, 3(3), 294-307.
- Muharemi, F., Logofătu, D., Andersson, C., and Leon, F. (2018). Approaches to building a detection model for water quality: a case study. In *Modern Approaches for Intelligent Information and Database Systems* (pp. 173-183). Springer, Cham.
- NHMRC, N. (2011). Australian drinking water guidelines paper 6 national water quality management strategy. *National Health and Medical Research Council, National Resource Management Ministerial Council, Commonwealth of Australia, Canberra*, 5-7.

- Qian, K., Jiang, J., Ding, Y. and Yang, S. (2020). Deep learning based anomaly detection in water distribution systems. In *2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1-6). IEEE.
- Ribeiro, V. H. A., Moritz, S., Rehbach, F., and Reynoso-Meza, G. (2020). A novel dynamic multi-criteria ensemble selection mechanism applied to drinking water quality anomaly detection. *Science of The Total Environment*, 749, 142368.
- Robles-Durazno, A., Moradpoor, N., McWhinnie, J., Russell, G. and Tan, Z. (2021). Newly Engineered Energy-based Features for Supervised Anomaly Detection in a Physical Model of a Water Supply System. *Ad Hoc Networks*, 102590.
- Sensus. (2013). Water 20/20: Bringing smart water networks into focus. *Sensus*. <http://sensus.com/smartwaternetworks> (Date of Access: Aug 5, 2021).
- Shalyga, D., Filonov, P. and Lavrentyev, A. (2018). Anomaly detection for water treatment system based on neural network with automatic architecture optimization. *arXiv preprint arXiv:1807.07282*.
- SPOTSeven Lab, 2018, “GECCO Challenge 2018 - Internet of Things: Online Anomaly Detection for Drinking Water Quality” (Date of access: Dec 06, 2021).
- SPOTSeven Lab, 2019, “GECCO Challenge 2019 – Internet of Things: Online Event Detection for Drinking Water Quality Control” (Date of access: Jan 02, 2021).
- Waterwise Rand water, 2017, Waterborne Diseases Factsheet, http://www.waterwise.co.za/site/water/diseases/water_borne.html (Date of access: Aug 10, 2021).
- Wu, P., Harris, C. A., Salavasidis, G., Lorenzo-Lopez, A., Kamarudzaman, I., Phillips, A. B. and Anderlini, E. (2021). Unsupervised anomaly detection for underwater gliders using generative adversarial networks. *Engineering Applications of Artificial Intelligence*, 104, 104379.

- Yang, Z., Liu, Y., Hou, D., Feng, T., Wei, Y., Zhang, J. and Zhang, G. (2014, October). Water quality event detection based on Multivariate empirical mode decomposition. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2663-2668). IEEE.
- Zhang, Y., Li, W., Zhou, P., Yang, J. and Shi, X. (2016, September). Big sensor data: a survey. In *International Conference on Internet and Distributed Computing Systems* (pp. 155-166). Springer, Cham.