MACHINE LEARNING BASED CONTAMINATION DETECTION IN WATER DISTRIBUTION SYSTEM

Akalewold Fikre [1*], Getachew Alemu [1]

[1] *School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia*

*****Corresponding Author Email address: akale98@gmail.com

DOI: https://doi.org/10.20372/zede.v42i.10185

ABSTRACT

Water is a necessary component of all human activities. According to the United Nations World Water Assessment Program, every day, 2 million tons of sewage, manufacturing, and agricultural waste are discharged into the world's water. Due to population demands and dwindling clean water supplies as well as available water pollution management mechanisms; there is an urgent need to use computational methods to intelligently manage available water. This paper proposes artificial neural networks, specifically, Convolutional Neural Networks, for automated water impurity detection. To refine the model, the picture of turbid water in the pipe was used to detect events. The algorithm of deep learning achieved 96.3 percent accuracy after extensive training with a dataset of 4220 images reflecting various levels of contamination. This shows that, the model can be used in water system pollution detection.

***Keywords:* -** CNN, NTU

1. INTRODUCTION

Drinking pure water is an issue for water supply companies around the world, and it is currently a well-known problem due to numerous vulnerable threats. An example is the 1993 contamination event in Milwaukee, which affected 403,000 people, resulting in thousands of hospitalizations and a hundreds of fatalities, with $96.2 million in medical and productivity costs [1]. Because of all the threats to public health, water system pollution detection is critical. In recent years, water quality sensors that are connected to the internet can be used to improve real-time monitoring of water quality. In the past, various approaches were proposed to address pollution detection issues, including single or multiple-type measurements that are analyzed separately or in combination from one or more locations in the network, using model-based or model-free approaches. The motivation of this work is the fact that, currently, machine learning techniques become promising for detecting contaminants in water quality. In this work we propose deep learning NN technique to determine the level of contamination of water based on an image.

Turbidity is a calculation of a liquid's relative visibility. Turbidity measures the concentration of light reflected by materials in the water; it increases when the materials found in the sample water rise. Clay, silt, organic and very tiny inorganic matter, algae, plankton, ingested colored organic compounds, and other microscopic species all contribute to the turbidity of water. Previous studies have found a relationship between turbidity levels and gastrointestinal disease [2]. More than 29 water quality parameter tests are monitored using a conventional method, including physical, chemical, and biological properties. From these

29 water quality parameters, in the operation that are most commonly and frequently sampled or monitored for water quality, include. temperature, disinfectant residual (chlorine concentration), pH, conductivity, ORP, and turbidity. Even among these selected parameters, the most frequently used water quality parameter to detect water contamination is chlorine concentration and turbidity. So from these two parameters, turbidity was chosen because the other water quality parameters directly or indirectly affect it.

The goal of this work was to develop a model-based approach for pollution detection throughout the water distribution system using turbid water images as they are closely correlated with the physicochemical properties of other water quality parameters. Researches done prior to our work have used different methods and obtained different results. Some studies [3, 4 and 5] planned three water quality detection methods based on a comparison of calculated and observed values and receiver operating characteristic (ROC) curves to test each technique. The US Environmental Protection Agency (EPA) has tested over 30 pollutants (including pesticides, insecticides, metals, and bacteria) that may be used in deliberate acts of water pollution [6, 7].

Yang et.al, [8] used a pilot-scale pipe system to monitor 11 pollutants at various concentrations using adaptive transformation of sensory measurements. As a result, pollutant classification and detection based on chlorine kinetics were made possible. The system developed by Guepie et.al, [9] was based on residual chlorine decay. Their hypothesis was that a contaminant in the WDS would absorb a large portion of the measured chlorine, and that this single parameter would provide enough information. The methods used in the previous studies were supervised classification methods. But when there are no real-time measurements of contamination events, the models must be trained and tested using simulated contamination events. In these models some random disturbances are added to the measured data to reflect the contaminant effect to preserve generality in the absence of sufficient information, In this regard, Eliades et.al, [10] looked into the issue of water quality by using a model- based method for detecting pollution events in water treatment systems. It takes into account well-known chlorine input signals and produces bounds of the predicted chlorine concentration at different sensing locations at each time stage by running several Monte-Carlo simulations in parallel with the real system. The result demonstrates that the system will adjust the detection bounds as the chlorine concentration input varies. Chlorination is the method of adding chlorine to drinkable water to disinfect it and destroy germs. It fixes one of many issues here. Contamination, on the other hand, occurs for a variety of causes and has varying effects on various parameters. As a result, detecting pollution based on chlorine concentration does not provide a complete control mechanism.

Mohammed et.al, [11] introduced Adaptive neuro- fuzzy inference system (ANFIS) models. This work prompted us to redesign the algorithm for certain water quality parameters that are highly correlated with one another and have an impact on the algorithm's efficiency.

It's intriguing to find out which parameters are closely correlated and then remodel the algorithm. In the work of Mohammad pour et al [12] used three separate algorithms, to investigate the issue of water quality. R2, RMSE, and MAE are used to compare results and they found SVM is competitive with neural networks in terms of results.

The system developed by Revathi, S. Kavi, and G. Shenbagalakshmi in [13] suggest using a wireless sensor network to develop and implement an actual system of water quality control for drinking water. The proposed system is low in cost, lightweight, and consumes little power.

An automated Aqua Sight water pollution detection method [14] uses a picture to assess the level of contamination in water. Here, they have used a convolutional neural network which involves an image of water to decide whether or not pollutants are present. In a report [18], it was suggested an intelligent real-time water quality monitoring strategy and concentrated on classifying water quality using machine learning techniques. The dataset includes dissolved oxygen, pH, conductivity, nitrate, biochemical oxygen demand, fecal coliform, and total coliform. In this study, among the different machine learning classifiers used, CAT Boost was determined to be the top classifier and stacking model. Shams et.al, [19] the grid search approach is applied in this research to adjust the parameters of four classification models and four regression models. It is reported that, for predicting WQI values, the MLP regressor model using the grid search strategy produced the best results, with an R2 equal to 99.8%.

Many aspects of the water quality at a specific place and time are covered by the Water Quality Index (WQI). The WQI computation is time-consuming and often influenced by errors when doing subindex computations. Therefore, developing an effective WQI forecasting method is essential.

2. MATERIALS AND METHODS

2.1. Collection of water samples

The first step in data processing is to obtain water samples from various locations. From various locations, a total of 216 water samples were collected. The samples were collected throughout different phases of a treatment plant, including raw water (Figure 1) during sedimentation, and finally from the reservoir (Figure 2) city reservoir, and household. Parts of the city where water pipes were exposed to domestic waste were chosen as sampling sites. Another criteria used in the selection process was proximity to contamination sources like industrial wastewater and hospital effluent outlets. Since the water pipe could not be accessed anywhere, locations had to be selected where it would be convenient to take samples if it broke down due to construction or some other reason. The majority of the samples were taken in the morning. The collection protocol followed the Addis Ababa Water and Sewage Authority's procedures (AAWSA).

Figure 1 The raw water at the time of sampling

Figure 2 Water Reservoir at the time of sampling.

The samples were analyzed for turbidity, pH, EC, TDS, Total Alkalinity, Calcium Hardness, Total Hardness, Magnesium Hardness, Ammonia, Nitrate, Nitrite, Phosphate, Fluoride, Iron, Manganese, Silica, Chloride, and Bicarbonate Alkalinity. All measurements were carried out in line with WHO requirements. Designer’s should consider changes throughout the physicochemical properties of water, concentrating on the turbidity of the water that is influenced by the chemical and biological particles of the contaminant (Figure 3 and 4), based on the findings of the previous parts. A single or multiple parameters may affect the quality of water. In the conventional method, if multiple parameters contribute to water pollution, the parameters and their contribution levels are identified in order to apply the treatment mechanism accordingly. For example, the most commonly known water quality parameters that appear together are iron and manganese. Sometimes, they both contribute to water pollution in different levels, to determine the behavior of water quality. This means the highly contributing parameters dominate in determining whether water is potable or not. But the treatment mechanisms consider both of the parameters, based on the WHO standard. In this research, machine learning determines the behavior of the water based on the turbidity of the water, whether a single or multiple parameters contribute to water pollution. We studied the effects of different parameters on the characteristics of turbidity. So if the turbidity values are higher than the normal values (5 NTU) due to a single or multiple parameters, the model will identify the water as potable. The dataset for the proposed method is generated based on the properties of the turbidity of the water. As the turbidity of the water rises, it decreases the purity of the water and the transmitted light by scattering and adsorbing the light.

Figure 3 Clean water with normal light

Figure 4 Contaminated water with 100mg/liter of sodium nitrate

When we look at clean water, it transmits light (see Figure 3). Clean water has normal light, and if it is contaminated, it decreases the transmission. (See Figure 4 contaminated water with 100mg/liter of sodium nitrate in normal light) For example, in the sample taken on January 8, 2020, total dissolved solids (TDS) for clean water were 74 mg/l, and in the sample taken the same day, total dissolved solids (TDS) for polluted water were 524 mg/l. So for the contaminated one, the water cloudiness increases greatly. The RGB of the images changes as the cloudiness of the images increases. This characteristic of the image plays a great role in determining if the water is potable or not.

2.2. Pre-processing

Noise can seriously affect the quality of digital images. Different factors may be responsible for introduction of noise in the image. In this phase of the system, we apply three different filters used for smoothing, sharpening and denoising to pre-process image in order to obtain a picture with more stable region. Eliminating the noise without blurring the details too much and enhancing edges without amplifying noise is very difficult. So, when using more than one filter, special care should be taken in order to make sure their effect is important. In this regard, the following filters were used sequentially.

i. Bilateral filter: To smooth the image
ii. High-pass filter: To sharpen images
iii. Median filter: To filter out noise from images

2.3. Model Design

Machine Learning (ML) algorithms are well-known for learning the underlying relationship in data and making decisions without the need for explicit instructions. Convolutional neural networks (CNNs) are a form of neural network that are used in deep learning. CNNs are large networks of nodes called neurons that create connections as they learn from data. Since CNNs require a person to identify specific features for a model to examine, they perform supervised learning. CNNs are one of the most powerful learning algorithms for comprehending image information, with excellent results in image segmentation, classification, identification, and retrieval tasks [15, 16].

CNN is divided into several learning levels, each of which consists of a mixture of convolutional layers, nonlinear processing units, and subsampling layers [17].

As a result, we constructed the model's structure, spawning numerous layers that perform various functions and contribute to the model's output in various ways. The convolution operation aids in the extraction of useful features (Figure 6) from data points that are globally correlated.

The non- linear processing unit (activation function) receives the output of the convolutional kernels, which not only aids in learning abstractions but also embeds non-linearity in the feature space. This non-linearity results in various activation patterns of different reactions, allowing it easier to understand semantic differences in images.

The proposed architecture consists of pooling and convolution layers alternated with one or more completely connected layers at the end. The convolutional layer executes a process called "convolution." Every neuron performs as a kernel throughout the convolution layers, which would be made up of a collection of convolution kernels.

Create the model: The model is made up of four convolution blocks, each with a max pool layer (Table 1). A relu activation feature is used to enable a completely connected layer with 512 units on top. This model hasn't been fine-tuned for high precision.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 73, 73, 64)	1792
max_pooling2d (MaxPooling2D)	(None, 36, 36, 64)	0
conv2d_1 (Conv2D)	(None, 34, 34, 64)	36928
max_pooling2d_1 (MaxPooling2	(None, 17, 17, 64)	0
conv2d_2 (Conv2D)	(None, 15, 15, 64)	36928
max_pooling2d_2 (MaxPooling2	(None, 7, 7, 64)	0
conv2d_3 (Conv2D)	(None, 5, 5, 64)	36928
max_pooling2d_3 (MaxPooling2	(None, 2, 2, 64)	0
flatten (Flatten)	(None, 256)	0
dense (Dense)	(None, 512)	131584
dense_1 (Dense)	(None, 5)	2565

Total params: 246,725
Trainable params: 246,725
Non-trainable params: 0

Table 1 The designed model.

The convolutional kernel divides the image into small slices, known as receptive fields, as shown in Figure 6. Extracting feature motifs is easier when an image is divided into small blocks.

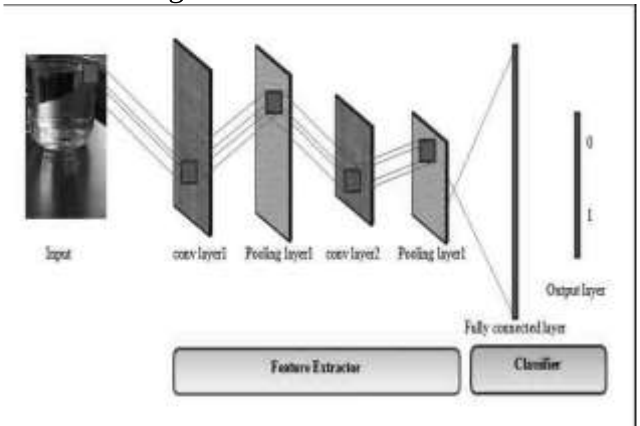

Figure 5 Architecture of the Convolutional Neural Network.

Once features have also been retrieved, their actual position is less important as long as their relative position to others is preserved. Down-sampling, also known as pooling (Figure 7) is an intriguing local process. It compiles similar data in the general neighborhood of the receptive field and outputs the dominant response for that area.

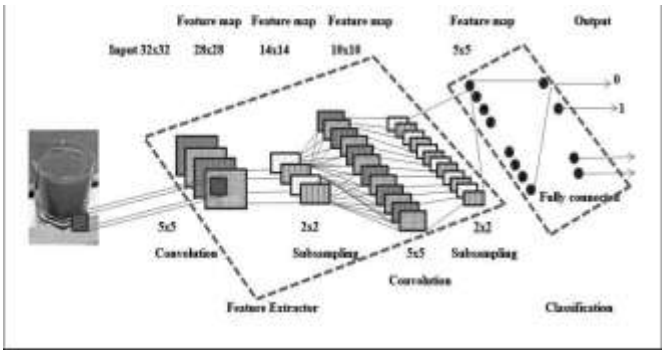

Figure 6 Segmentation Process.

Max pooling takes the largest element from the rectified feature map (Figure 6). Taking the largest element could also take the average pool.

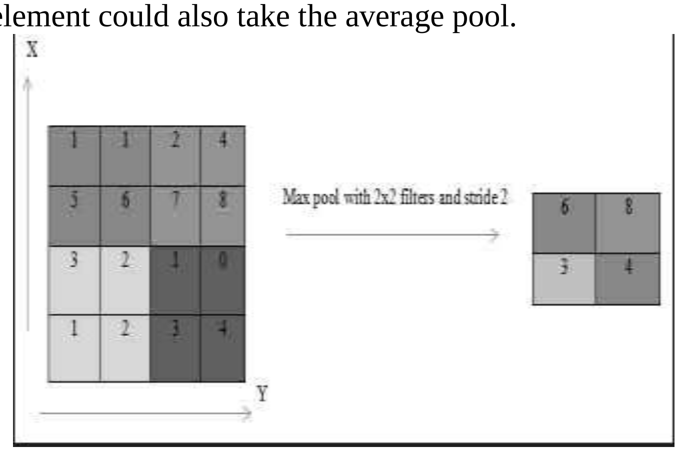

Figure 7 Max Pooling.

3. RESULTS AND DISCUSSIONS

The system model error rates are significant indicators in assessing the model's success in the proposed architecture. The accuracy indicates the likelihood that the picture will fit the target mark correctly. The accuracy level of the model varies throughout the training cycle due to the dataset used for training and validating the model. The model which was created in the previous section is made up of 4 convolutional layers, 4 max pooling layers, and 1 dense layer. The accuracy of validation was 80.49%. This means we can correctly classify 80.49% of the images in the validation collection that the model missed.

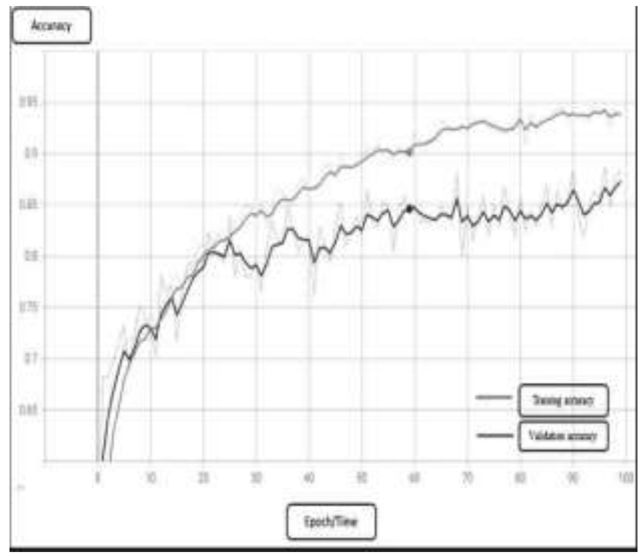

Figure 8 Training and validation accuracy for the proposed model

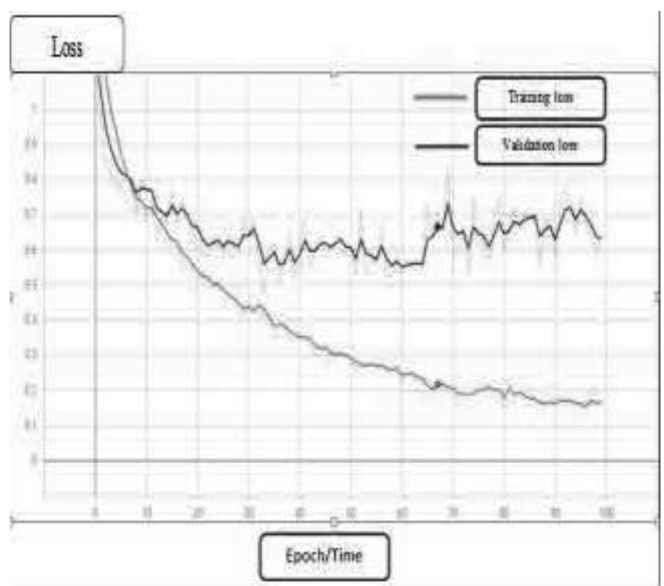

Figure 9 Training and validation loss for the proposed model

From Figures 8 and 9, we can see that the training accuracy is above the validation accuracy and training loss is way below the validation loss, especially after the 20th epoch. When there are few training datasets, the model can learn from sounds or unwanted information in the training dataset, which can have a negative effect on the model's output on the new dataset.

This phenomenon is known as over-fitting. It means that the model will have a difficult time generalizing on the new dataset. There are many methods for combating over-fitting during the training period. Using data augmentation and adding Dropout to the model are the two main methods.

Data augmentation is a technique to produce further training data from an existing dataset by augmenting it with random transformations that result in believable-looking photos. This allows the model to be exposed to more facets of the data and generalize more effectively. The updated model was trained with 4220 images and its testing dataset contained 1055 images, based on the principle of data augmentation.

Dropout, the kind of regularization, is another technique for reducing over-fitting in the network. When Dropout is applied to a layer, it randomly removes a number of output units from the layer during the training phase (by setting the activation to zero). Dropout accepts fractional numbers as in the manner of 0.1, 0.2, 0.4, and so on as data.

This means that 10%, 20%, or 40% of its output nodes from the applied layer would be randomly removed. In addition to that, it helps to optimize the model until it achieves the best accuracy on training and validation, as well as to get relatively low loss model architecture.

Layers and nodes per layer, as well as 0, 1, or 2 dense layers, are the simplest things to change in the model. Finally, the model achieved an almost match with well-balanced training and validation metrics (Figures 10 and 11), with validation accuracy of 94.82 percent and the highest (lowest) validation loss of 0.1719. The result was achieved using two convolutional layers, each with 64 nodes, 2 max pooling layers, and one dense layer followed by a single dropout.

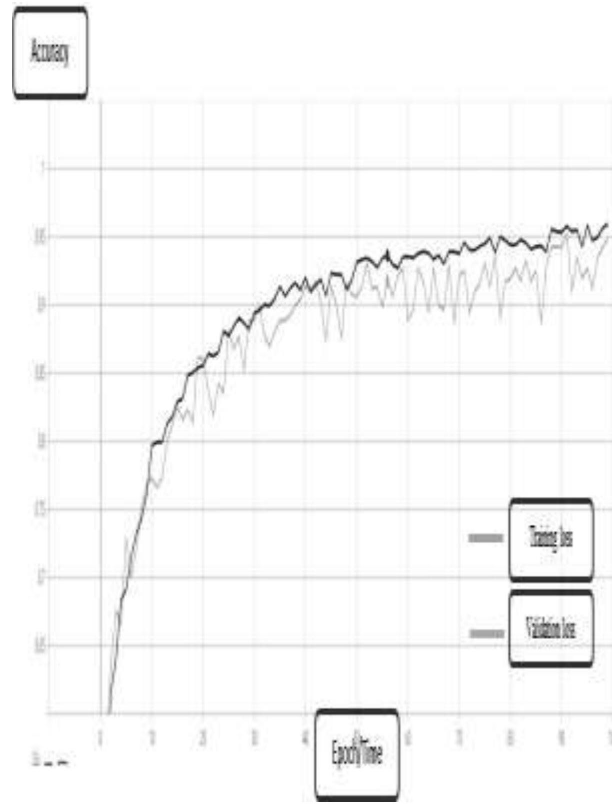

Figure 10 Training and validation accuracy for the best models

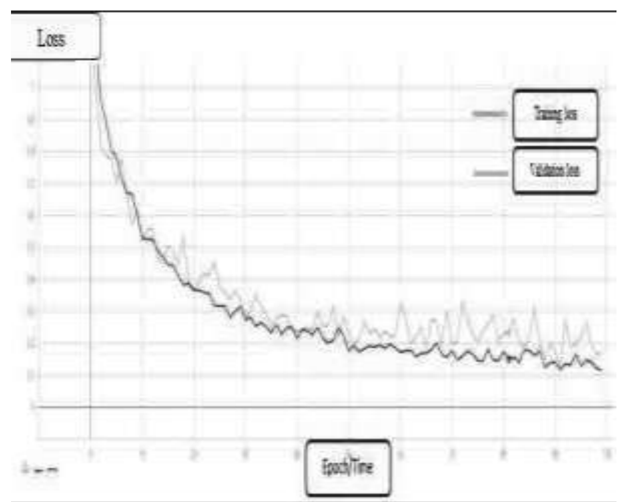

Figure 11 Training and validation loss for the best models

In addition to metrics directly from the model and its prediction values, F1-score, Precision, Sensitivity, and Accuracy can all be calculated. In general, classification accuracy can mask the information essential to diagnose the model's performance. As a result, generating a confusion matrix can help to figure out what the classification model is getting right and where it is going wrong. Table 2 is a multiclass confusion matrix of the classes.

Table 2 A multiclasses Confusion Matrix of the classes

	Predicted					
	Class	Chemical	Clay	Clean	Sand	Silt
Actual	Chemical	208	0	0	0	2
	Clay	0	206	0	5	1
	Clean	0	3	203	3	1
	Sand	1	1	5	201	4
	Silt	2	7	0	4	198

3. CONCLUSIONS

The system model error rates are significant indicators in assessing the model's success in the proposed architecture. The accuracy indicates the likelihood that the picture will fit the target mark correctly. The accuracy level of the model varies throughout the training cycle due to the dataset used for training and validating the model. The model was made up of 4 convolutional layers, 4 max pooling layers, and 1 dense layer is used. The accuracy of validation was 80.49%. This means we can correctly classify 80.49% of the images in the validation collection that the model missed.

4. CONFLCT OF INTEREST

The authors declare that there is no conflict of interest in this work.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Addis Ababa University, Addis Ababa Institute of Technology, School of Electrical and Computer Engineering. The accomplishment of the paper would not have been possible without its support

REFERENCES

[1]. Corso, P., Kramer, M., Blair, K., Addiss, D., Davis, J., Haddix, A., *"Cost of illness in the 1993 waterborne cryptosporidium outbreak, Milwaukee, Wisconsin,"* Emerging Infectious Diseases vol. 9. 2003, pp. 426–431.

[2]. Schwartz, J., Levin, R., *"Drinking water turbidity and health."* Epidemiology, vol. 10 (1). 1999, pp. 86-90.

[3]. Klise, K.A., McKenna, S.A., *"Multivariate application for detecting anomalous water quality."* In: Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium. WDSA, Cincinnati, Ohio, USA, 2006, pp. 1-11.

[4]. Klise, K.A., McKenna, S.A., *"Water quality change detection: multivariate algorithms"*. In: Saito, T.T., Lehrfeld, D. (Eds.), Proc. SPIE, vol. 6203. 2006, pp. 1–9.

[5]. McKenna, S.A., Wilson, M., Klise, K.A.,*"Detecting changes in water quality data."* Journal of the American Water Works Association, vol. 100 (1). 2008, pp. 74-85.

[6].https://nepis.epa.gov/Exe/ZyPDF.cgi/P1004B3M.PDF?Dockey=P1004B3M.PDF. (Accessed on: Jan 13, 2021).

[7].https://nepis.epa.gov/Exe/ZyPDF.cgi/P1004B4D.PDF?Dockey=P1004B4D.PDF. (Accessed on: Jan 13, 2021).

[8]. Yang, J.Y., Haught, C.R., Goodrich, A.J., *"Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors:"* techniques and experimental results. Journal of Environmental Management, vol. 90 (8). 2009, pp. 2494-

2506.

[9]. Guepie, B.K., Fillatre, L., Nikiforov, I., "*Sequential monitoring of water distribution network.*" In: Paper Presented at the IFAC Proceedings Volumes (IFAC-papers Online), vol. 45 (16). 2012, pp. 392-397.

[10]. Eliades, D. G., Lambrou, T. P., Panayiotou, C. G., & Polycarpou, M. M. *"Contamination event detection in water distribution systems using a model-based approach."* Procedia Engineering, vol. 89. 2014, pp. 1089–1096.

[11]. Mohammed, H., Hameed, I. A., & Seidu, R. "*Machine Learning – Based Detection of Water Contamination in Water Distribution Systems*." GECCO18. 2018. pp.1664–1671.

[12]. Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., Ab Ghani, A., & Chan, N. W., *"Prediction of water quality index in constructed wetlands using support vector machine."* Environmental Science and Pollution Research, vol. 22(8). 2015, pp. 6208–6219.

[13].Kavi Priya, S., Shenbagalakshmi, G., & Revathi, T., *"Design of smart sensors for real time drinking water quality monitoring and contamination detection in water distributed mains."* International Journal of Engineering & Technology, vol. 7(1.1) 2017, pp. 47-51.

[14]. https://arxiv.org/pdf/1907.07573.pdf (Accessed on: May 30, 2020).

[15]. Ciresan, D., Giusti, A., Gambardella, LM. Schmidhuber, J., *"Deep neural networks segment neuronal membranes in electron microscopy images."* In: Advances in neural information processing systems. 2012. pp 2843–2851

[16]. Liu, X., Deng, Z., Yang, Y., *"Recent Progress in semantic image segmentation."* vol. 52. 2019, pp. 1089-1106.

[17]. Xiang, Y., & Jiang, L. (2009). *"Water quality prediction using ls-svm and particle swarm optimization."* Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on (pp. 900–904). IEEE.

[18].Nida Nasir, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, Ahmed Al-Shamma'a, *"Water quality classification using machine learning algorithms"*, Journal of Water Process Engineering, vol. 48, 2022, 102920, ISSN 2214-7144,

[19]. Shams, M.Y., Elshewey, A.M., El-kenawy, ES.M. et al. *"Water quality prediction using machine learning models based on grid search method"*. Multimedia Tools Appl (2023).