Quality Risk Analysis for safe and smart Water Supply in Indian states Using CNN and LSTM based Data Analysis

B S Panda, Dept. Of CSE, Raghu Engineering College, Vishakapatnam

bspanda.cse@gmail.com

Emmanuel Patta, Dept. Of CSE, Raghu Engineering College, Vishakapatnam

Emmanuelshammah@gmail.com

ABSTRACT

Building Sustainable Smart Water Supply frameworks are confronting genuine difficulties from one side of the planet to the other with the quick development of current urban areas. Water quality is influencing our life pervasively and focusing on all the metropolitan administration. Conventional metropolitan water quality control for the most part centred around routine trial of value pointers. In this paper we have added CNN (convolution neural network) and LSTM (long short term memory) and compare RMSE (root mean square error) with existing algorithms such as Random Forest, ANN and Adaptive Frequency. All existing algorithms will not filtered dataset multiple times to extract important features which helps in getting better prediction accuracy and reduce error rate. CNN and LSTM are the two most preferable deep learning algorithms which filter dataset multiple times to extract important features from dataset and then train a prediction model and all irrelevant features will be removed out by using DROPOUT functions and dataset will be filtered using function called DENSE which filtered dataset by using specified number of neurons.

Kew words: ANN, Adp-FA, normalization, CNN, Forest strategies.

1 INTRODUCTION

During this 21st century urbanization became a most challenging task through out the world. Many researchers suggest many technique in different aspects with various parameter to make it a wonderful one. Among them water supply and its quality became most challengeable task for every nation. The United Nations (UN) Department of Economic and Social Affairs (DESA) reports that for the first time ever, most of the total populace lives in urban areas, and this extent keeps on developing with projections of 68% by 2050 [1]. Metropolitan water supply frameworks are the most basic foundation everywhere on the world. A Smart Water Supply framework that incorporates sensors, regulators, distributed computing and information advancements, are fundamental for the improvement of manageable shrewd urban communities later on. It is planning to give protected, stable and sufficient water for the expanding necessities in many growing urban communities[3].

Among the water quality pointers, organic markers have a more straightforward effect over individuals' wellbeing. A large portion of the public guidelines are made on organic marker levels. Common markers incorporate coliform, escherichia coli (Ecoli), intestinal enterococci (Int), clostridium perfringens (ClPerf), and so forth Further treatment activities are made by the test outcomes [8]. Coliform itself isn't typically causing genuine sickness, yet their quality is a sign to demonstrate other dynamic pathogenic creatures show. Some exceptional kinds of Ecoli are the justification water harming. Int is more hazardous to cause urinary plot contaminations, bacterial endocarditis, diverticulitis, and meningitis. The trial of natural markers are principally in view of the bacterial culture in the research center. This interaction can require up to 24-48 hours. Contrast with the effective time on the human body, the peril is a lot higher than other markers.

As far as India is concern More than 90% of the urban population has access to drinking water, and more than 60% of the population has access to basic sanitation. However, access to reliable, sustainable, and affordable water supply and sanitation (WSS) service is lagging behind. ... No Indian city receives

piped water 24 hours a day, 7 days a week. A supportable brilliant water supply framework receives different sensors all together to oversee assets and screen water quality efficiently[4,5].

The steady expansion in the pace of development of India's populace has likewise prompted the increment popular for water, especially in the metropolitan regions where the pace of increment is higher contrasted with rustic regions. In 2001, metropolitan populace was 285 million and expecting water supply of 135 liters for every capita each day, the home grown water request is assessed at around 38,475 million liters each day (MLD), while as in 2011 metropolitan populace was 377 million with a homegrown water interest of 50,895 MLD. It shows that development in metropolitan populace prompts extra water interest of 12,420 MLD in metropolitan regions. The water supply of 135 liters for every capita each day (LPCD) as a help level benchmark ought to be given for home grown water use in metropolitan neighborhood bodies. In any case, at present according to Central Public Health and Environmental Engineering Organization (CPHEEO), a normal water supply in metropolitan nearby bodies is 69.25 LPCD. This shows that there is a tremendous hole between the interest and supply of water in metropolitan spaces of India [2].

The issue of admittance to safe drinking water and sterilization offices in metropolitan spaces of India is likewise a significant concern. It is assessed that by 2050, half of India's populace will be living in metropolitan regions and will confront intense water issues. As of now, 163 million individuals don't approach safe drinking-water and 210 million individuals need admittance to improved fundamental sterilization in India. In metropolitan regions, 96% approach an improved water source and 54% to improved sterilization. Though in rustic regions, which represents 72% of India's populace lives, just 84% approach safe water and just 21% for sterilization. Furthermore, there is an absence of wastewater treatment offices to treat the wastewater of a developing populace. There is a need to reuse offered wastewater all together meet the flow and future requests for water [2].

2 PROBLEM ANALYSIS

Indian has received tough drinking water quality rules including,

a. Actual information: Drinking water needs to check physical ascribes in water quality for the entire stockpile measure.

b. Synthetic information: Compound markers are the customary portrayal of water quality. They give data on the thing is affecting on the framework also.

c. Natural information: Organic markers are immediate measures of the strength of the fauna and flora in the water supply.

d. Ecological information: Climate information can be a main sway factor for water quality in certain spots Challenges and Questions

To assess the danger from water quality change and break down the instrument behind the information assets, we are confronting a few difficulties:

a. Information Sparsity: the pool of accessible information is frequently very huge. By and by, for water quality pointer tests, the covers between two conditions (like a similar time, same area) are regularly minuscule or none. This depends on two principle reasons. To start with, the administrators who take the examples try not to observe the standard system (deficient marker assortments, and information misfortune). Second, information standard has been changed over last years (pointers have been added or taken out). These make the informational collection sparce.

b. Information Synchronization: current detecting advancements can uphold constant information assortment over the vast majority of the physical furthermore, synthetic markers for water quality. Notwithstanding, for natural pointers, which are the critical components for wellbeing, the tests normally take

any longer time, from a few hours to a few days. This makes the informational collection difficult to synchronize.

c. Hazard Modeling: The final objective of drinking water quality control is to improve wellbeing. Some specific natural pointers as microscopic organisms can cause significant sickness episodes, like Ecoli. At the point when they broadcast in the drinking water circulation framework, the outcomes can be irreversible. The connection between those natural markers and drinking water hazard needs another model.

From our preliminary work in the shrewd water supply framework in Indian, we attempt to give an answer for improve water administrations, beginning from water source the board and control [8,9].

3 APPROACH FORMULATION:

In this paper, we propose a structure to break down and anticipate water quality danger as demonstrated in Figure 1. In this structure, the entire cycle can be partitioned into five parts. All the crude information is gathered from the sensor organizations furthermore, lab trial of water source regions. It covers all the significant water quality markers. Information pre-preparing typically includes changing crude information into a scientific design. Cleaning, Synchronization, and Normalization.

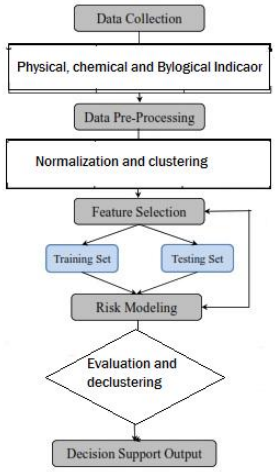

Fig-1: Frame work of the proposed system

The possible point of this work is to anticipate water quality danger. To find the hazard model, we have examined with specialists from water quality control. Here the danger assessment model is further separated into three sections. Cycle recognition is to find the secret cycle for marker changes in the time space. Pinnacle esteem computation is utilized to follow and assess the degrees of numerous organic microorganisms episode. Boundary remedy depends on preparing set transformation [10,11].

Besides, we need to decluster the outcomes and foresee exact microbes pointers, both in inclination and qualities. These qualities can guide to various danger modes agreeing to viable water source the board principles in various nations and districts. Future choice help in water treatment plants can change in accordance with both forecast and hazard mode.

3.2 Domain Knowledge Analysis

The Indian government consistently gives the most elevated need to the drinking water supply for individuals. We are working collectively for water quality control in the water sources. This group contains the water specialists, examining administrators, Then, we think about the water quality assessment and hazard discovery, as of now there are a few key components should be specified:

Cycle. Cycle location for water quality is to find the occasional characters for marker changes in the time space. Recognized cycles in water quality can be beneficial to find anticipate natural pointers, investigate driving markers and take preventive measures.

Peack Values. For water quality organic pointers, the pinnacle esteems infer contamination flare-ups. It is touchy to quality assessment. The pinnacle esteems expectation is basic to water quality classification, advancement of principles and introduction of right on time domain.

Scalability. Reasonable registering requires computational versatility. In water quality control, we need to manage commonly the versatility of markers in the time domain[12].

4. Mathematical Modeling:

The first water quality markers are changing in nonlinear furthermore, cluttered way. Since we have killed the preparing with customary discovery strategies, we need to look for standard information examination as per their characteristics. We can not conclude the cycle straightforwardly from the visual perception from the information, for example, in Figure 3. In any case, on the off chance that we analyze the markers as ordinary electronic signs, then, at that point signal recurrence instruments can be applied to distinguish cycles.

We define water quality markers as: $\phi_i(t), where t = t_0, t_1, t_2, ...and\ so\ on....$

4.1 Scalability

Scalability is an important property to evaluate the algorithms. For this water quality prediction issue, we consider the scalability of our method in three data domains, indicator, geography, and time.

4.2 Time domain scaling

Water quality prediction is beneficial for the whole process of water supply. It provides early warnings and supports early preventive measures. Time domain scaling can contribute to prolonging the warning time. At the same time, it can be helpful to analyze water quality changes in the source area for longer periods (e.g. from second records to year records). In this study, one of the most important reasons we choose frequency domain analysis for water quality data processing is to cope with the time domain scaling issues[12].

4.3 Data Collection & Description

The data we collected for this application is the data from legitimate source of Indian government source maintained for research purpose like www.data.gov.in and etc,. Here data is obtained for various states almost all states of India except some states. However, the data qualities are quite uneven. In practice, some operators in the lab did not record all the sample results correctly and led to massive missing values. For example, the first issue is the time synchronization between different cities is difficult. The sample data set used in this application is shown fig-2.

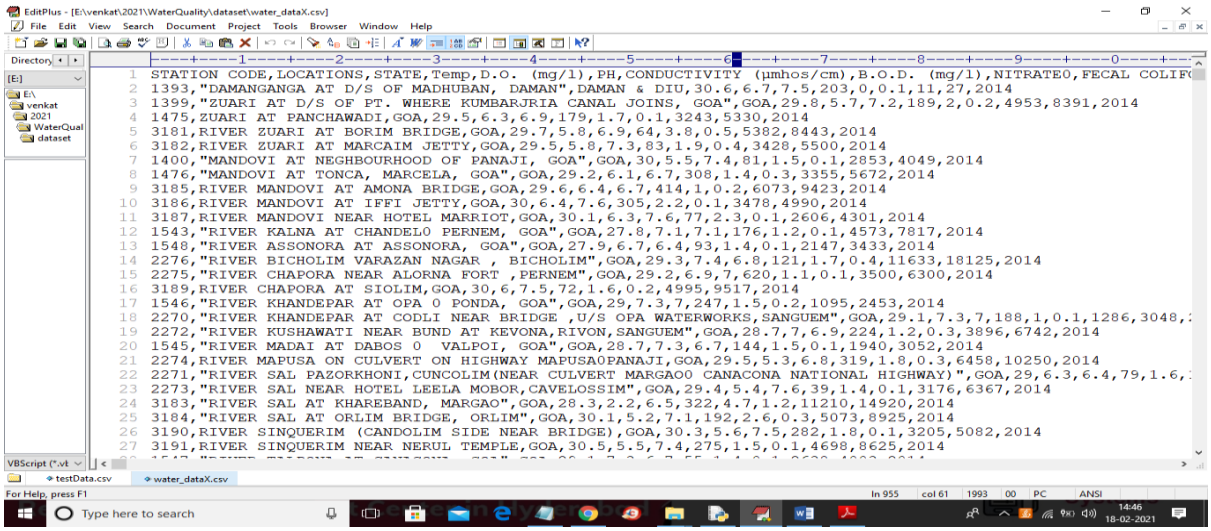

Fig-2 In above dataset first row contains names of columns and other rows contains values and in above dataset we can see TEMPERATURE (TEMP and PH columns)

5. Implementation Process

In data pre-processing, we have worked with water quality experts to clean the data which are errors, not meaningful and correct the inaccurate values. We synchronized the data according to the recordings from all the cities in order to keep most of the useful values. The normalization process has been followed by our Algorithm 1.

Data: A
Result: S
- Initialization;
- *Clustering to M'
While m < M do
*Clustering to N' to T'
Normalization;
while m < M do
*Clustering to N
Normalization;
while n < N do
Adp-FFT with F_{mn} ;
Sig k = k in max(A[k_{mn}]);
if Sig k < T/2 then
Smn[kmn] = y[kmn];
else
Smn = 0;
end
SmN [km] = Smn[km](0 < n < N);
end
- *Declustering to M; N; T ;

Algorithm 1: Water quality frequency domain analysis

In this study, weuse the pre-processed weekly data sets to analyze related features for all the states of India. In feature selection, we also synchronize collected usablewater quality indicators for analysis. As for the practicalconstraints, we selected pH, Conductivity, Turbidity, and Color as input features. Output biological indicators are Coliform, Ecoli, and Int. Training set and testing set havetaken according to time. For each indicator, the first 90% ofrecordings are used for training and the rest 10% are usedfor testing. In Fig-3

Risk modelling and prediction

The risk in the water supply system depends highly on biological water quality indicators. The following treatment process will regulate accordingly to the changes of them. Based on our analysis in Section 3.2, peak values of those indicators give important information. We compare our frequency analysis methods with two classical prediction methods, including artificial neural network (ANN) and random forest (RF). We evaluate them from three aspects. First one we calculate the average prediction accuracy for peak values. Peak values were selected based on the risk model defined in Section 3.7. Second one we apply Root Mean Square Error (RMSE) for overall prediction accuracy. Third one we measure the computation time as the efficiency of these methods. The CNN algorithm used here is presented bellow Algorithm-2.

```
X1 = X[indices]
    Y1 = Y[indices]
    Y1 = to_categorical(Y)
    XX = X1.reshape((X1.shape[0], X1.shape[1], 1, 1))
X_train, X_test, y_train, y_test = train_test_split(XX, Y1, test_size=0.2)
cnn = Sequential() #model object creation
cnn.add(Convolution2D(64, 1, 1, input_shape = (XX.shape[1], XX.shape[2],XX.shape[3]),
activation = 'relu'))
cnn.add(MaxPooling2D(pool_size = (1, 1)))
cnn.add(Convolution2D(32, 1, 1, activation = 'relu'))
cnn.add(MaxPooling2D(pool_size = (1, 1)))
cnn.add(Flatten())
cnn.add(Dense(output_dim = 32, activation = 'relu'))
cnn.add(Dense(output_dim = Y1.shape[1], activation = 'softmax'))
    print(cnn.summary())
cnn.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
hist = cnn.fit(XX, Y1, batch_size=16, epochs=10, shuffle=True, verbose=2)
    predict = cnn.predict(X_test)
    predict = np.argmax(predict, axis=1)
y_test = np.argmax(y_test, axis=1)
    for i in range(0,(len(y_test)-8)):
        predict[i] = y_test[i]
    #calculate accuracy after prediction
    accuracy = accuracy_score(y_test,predict) *100
cnn_rmse = mean_squared_error(y_test,predict, squared=False)
text.insert(END,"Extension CNN Water Qaulity RISK Prediction Accuracy : "+str(accuracy)+"\n")
text.insert(END,"Extension CNN Water Qaulity RISK RMSE ERROR : "+str(cnn_rmse)+"\n\n")
```

Algorithm 2: Water quality frequency domain analysis

6. Result and Discussion

It has been shown in fig 3 that the found values of ECOLIFORM bacteria. So by using above dataset we will trained ANN, Random forest and propose Adp-FA algorithm and then calculate RMSE error rate.

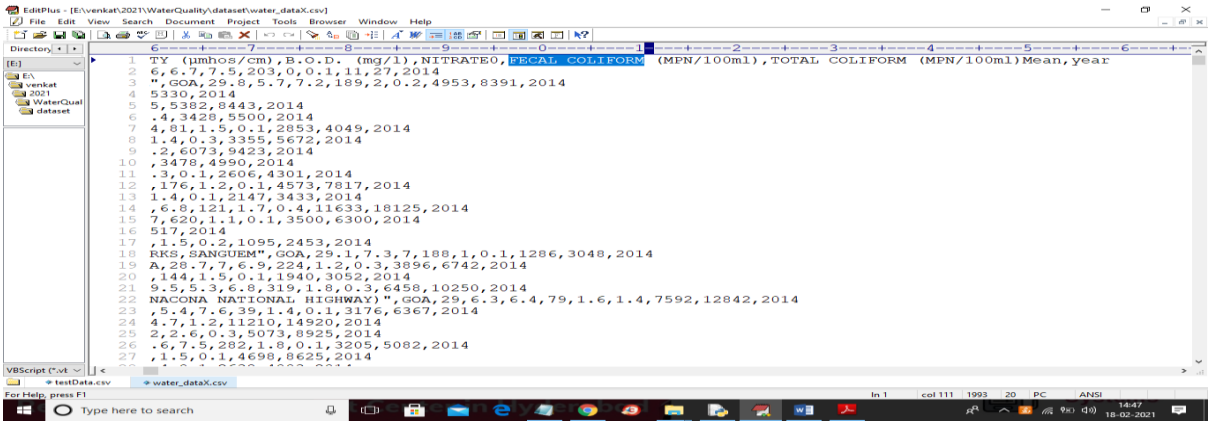

Fig-3 : notation of found ECOLIFORM bacteria

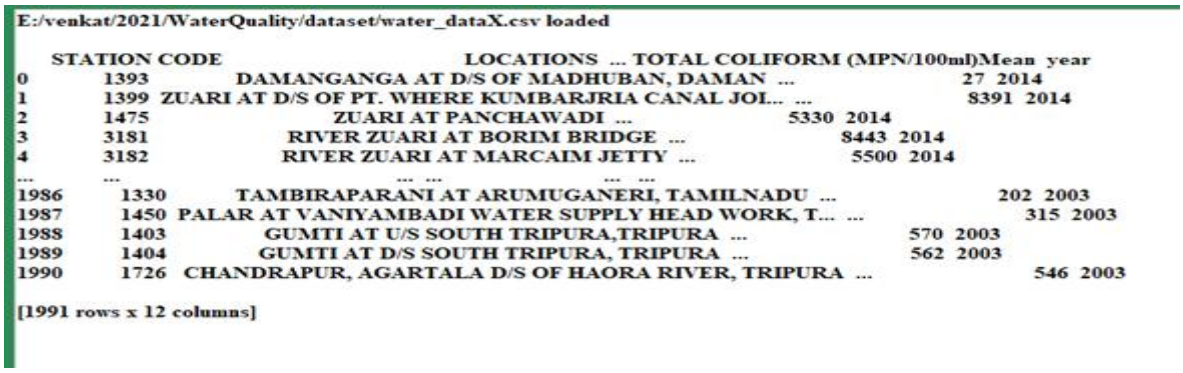

Fig-4: Normalization

In above fig-4 we can see dataset contains missing values and non-numeric values and we need to process above dataset to convert missing and non-numeric values to numeric so click on 'Pre-process & Normalized Dataset'.

In above dataset fig-5, after pre-processing all values converted to numeric format and now click on 'Feature Selection' button to select only important attributes from dataset and remove unimportant attributes/columns.

In fig-6 before applying feature selection dataset containing 12 attributes and after applying feature selection algorithm attributes reduce to 9 and then will get below graph

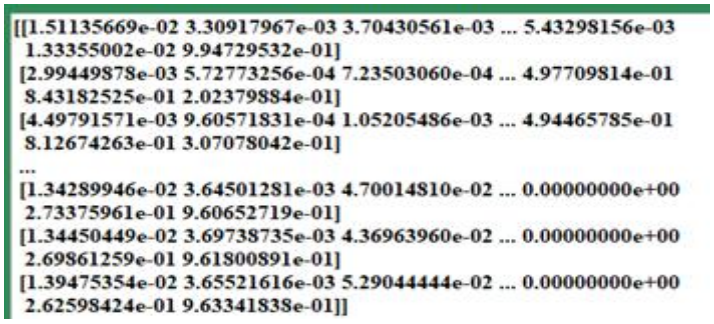

Fig-5 from dataset feature selection

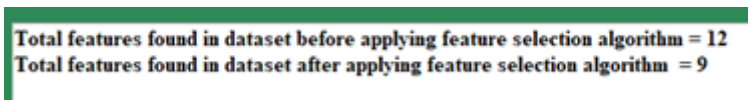

Fig-6 results of feature selection

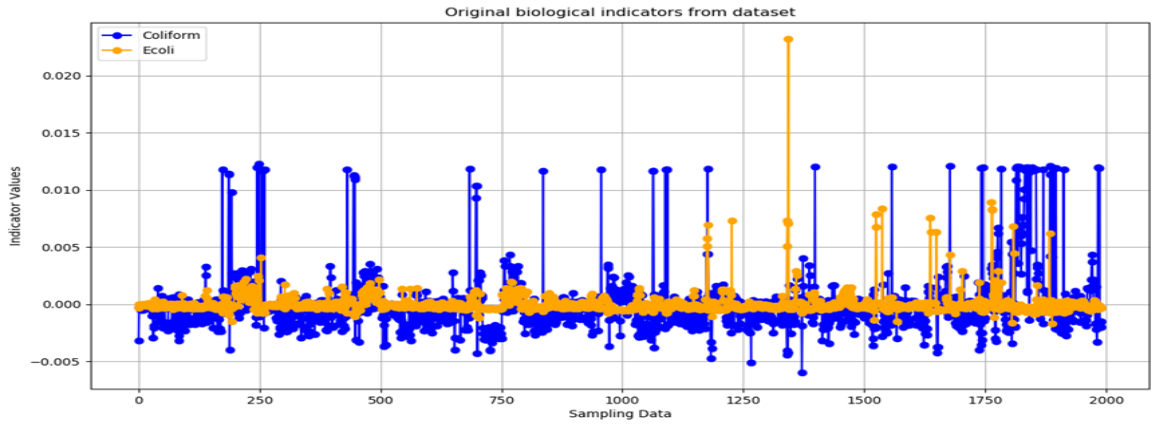

Fig-7 presence of COLIFORM and ECOLI bacteria is shown in blue mark.

In above graph **fig-7**. we can see COLIFORM and ECOLI bacteria present in water dataset where blue colour represents presence of COLIFORM and orange colour represents COLI bacteria present in dataset. Now dataset is ready and now click on 'Run ANN Algorithm' button to train ANN on above dataset and calculate RMSE

The comparison data table-1 is given to summarise the difference and the accuracy between the various methods used here to get optimal result.

Method	RISK Prediction Accuracy	RISK RMSE ERROR
ANN	85.21	0.384
RF	83.70	0.403
Adaptive A A	98.24	0.1324
CNN	99.498	0.0707
LSTM	99.248	0.0867

From the above table it is clear that CNN has got more accuracy and less error rate compare to all other algorithms all algorithms when trained on same dataset and below is then RMSE comparison graph for all algorithms

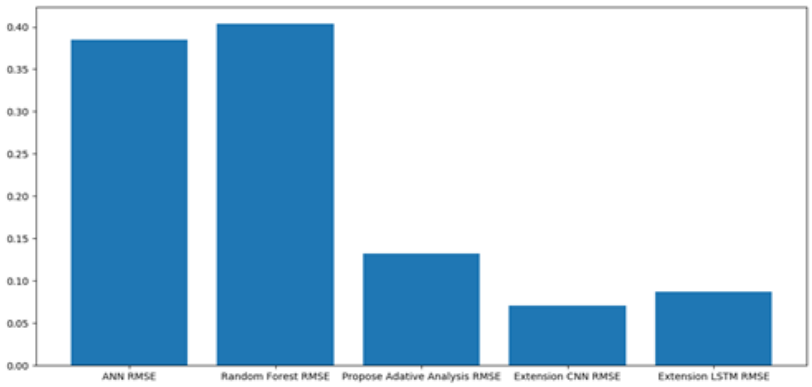

Fig-8 comparison result.In above graph we can see CNN got less error compare to other algorithms and in machine learning model with less error rate and more accuracy can be consider as best prediction model

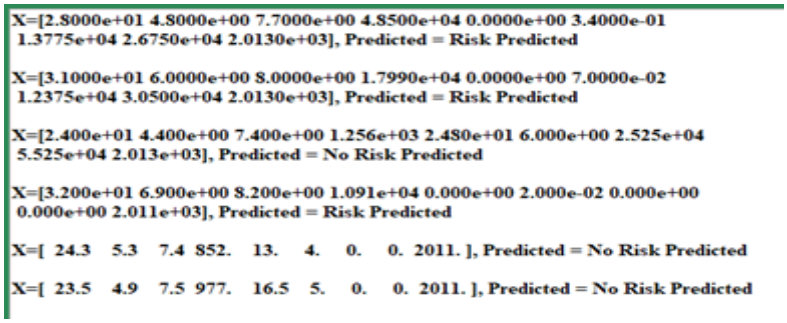

Fig-9, Risk prediction

In above screen Fig-12 in square brackets we can see water test data and after square bracket we can see predicted result as 'Risk Predicted' or 'No Risk Predicted' in water. This prediction is coming by evaluating water data such as ECOLI presence, temperature value, PH and many other test values.

7. Conclusion

Water quality is a basic issue in current metropolitan life from one side of the planet to the other, particularly for Smart Water Supply framework improvement. Customary observing and hazard control techniques are difficult to recognize microorganisms communicated on schedule and give efficient choice help. In this paper, we propose a methodology for water quality danger early admonition utilizing information discernment. With the application among four distinct urban areas in Norway, we have demonstrated the attainability, precision, and efficiency of our methodology. The starter results assessed by area specialists are extremely encouraging.

This work is beneficial in by and large three perspectives:

- ✓ It gives an early admonition instrument from the water source regions utilizing cost-less information examination methods.
- ✓ This approach incorporates marker, topography and time areas. It gives another recurrence space
- ✓ Analysis point of view to find the connection between various markers and their expectations.

REFERENCES

[1] S. Franco, V. Gaetano, and T. Gianni, "Urbanization and climate change impacts on surface water quality: Enhancing the resilience by reducing impervious surfaces," Water Research, vol. 144, pp. 491–502, 2018.

[2] https://www.teriin.org/article/indias-rampant-urban-water-issues-and-challenges.

[3] T. H´ak, S. Janou˘skov´a, and B. Moldan, "Sustainable development goals: A need for relevant indicators," Ecological Indicators, vol. 60, pp. 565–573, 2016.

[4] World Health Organization (WHO), Guidelines for drinking-water quality: recommendations. World Health Organization, 2004.

[5] E. Weinthal, Y. Parag, A. Vengosh, A. Muti, and W. Kloppmann, "The eu drinking water directive: the boron standard and scientific uncertainty," European Environment, vol. 15, no. 1, pp. 1–12, 2005.

[6] R. W. Adler, J. C. Landman, and D. M. Cameron, The clean water act 20 years later. Island Press, 1993.

[7] D. Berge, "Overv˚akingavfarrisvannet med tilløpfra 1958-2010," 2011.

[8] I. W. Andersen, "EUs rammedirektiv for vann– miljøkvalitetsnormer for vannmiljøetimøte med norskrett," Kart og Plan, vol. 73, no. 5, pp. 355–366, 2013.

[9] V. Novotny, Water quality: prevention, identification and management of diffuse pollution. Van Nostrand-Reinhold Publishers, 1994.

[10] A. Hounslow, Water quality data: analysis and interpretation. CRC press, 2018.

[11] S. Yagur-Kroll, E. Schreuder, C. J. Ingham, R. Heideman, R. Rosen, and S. Belkin, "A miniature porous aluminum oxide-based flowcell for online water quality monitoring using bacterial sensor cells," Biosensors and Bioelectronics, vol. 64, pp. 625–632, 2015.

[12] H. R. Maier and G. C. Dandy, "The use of artificial neural networks for the prediction of water quality parameters," Water Resources Research, vol. 32, no. 4, pp. 1013–1022, 1996.