Water Quality Monitoring And Forecasting UsingMachine Learning

Bojja Navi Namratta
Electronics and Communication department
Sreenidhi institute of science and technology
Hyderabad, India

Chenna Nikhitha
Electronics and Communication department
Sreenidhi institute of science and technology
Hyderabad, India

Deepak Menon
Electronics and Communication department
Sreenidhi institute of science and technology
Hyderabad, India

Dr. S Latha
Electronics and Communication department
Sreenidhi institute of science and technology
Hyderabad, India

Mr. Y. Sreenivasulu
Electronics and Communication department
Sreenidhi institute of science and technology
Hyderabad, India

***Abstract*— Water is a crucial resource and has a wide range of applications. Water is essential for the sustainability of life on planet Earth. From bathing to cooking, water is used in almost every activity we perform on a daily basis.**

Earth is covered by almost 70% water but not all of it is useable and safe for utilization. Only about three percent of Earth's water is freshwater. Of that, only about 1-2 percent can be used as drinking water; the rest is locked up in glaciers, ice caps, and permafrost, or buried deep in the ground. Most of our drinking water comes from rivers and streams.

In our constantly developing world, the increased rate of pollution and urbanization have started to contaminate the water supply making it unfit for drinking. Thus, a water quality monitoring system is needed to monitor and forecast its quality, and safeguard human health. To perform this task a system is developed using the popular Python programming language which helps to calculate the quality of water based on the readings taken from various sensors. We use an Auto-Regressive Integrated Moving Average (ARIMA) model to monitor water quality.

Keywords— ***Python, Water quality forecasting, ARIMA model***

I. INTRODUCTION

Water used for our day-to-day activities are usually stored in overhead tanks. These overhead tanks can become breeding grounds for various microorganisms and pathogen making the water unfit for drinking and usage. Hence a pressing need arises for a reliable water quality monitoring system that helps to monitor and forecast it so that a healthy and hygienic livelihood can be facilitated.

Water contamination is a serious problem and should not be overlooked. Contaminated water is due to many reasons such as deposit of thermal and industrial wastes/chemicals into water bodies, throwing plastic in water and many more. We as humans should collectively reduce our plastic consumption so that water bodies can be safe. Along with this we have to find other sustainable methods of disposing thermal wastes. Contaminated water can cause a number of diseases such as cholera, typhoid, dysentery, guinea worm, hepatitis etc.

AI has become popular these days and its scope is only increasing every day. Various machine learning and AI algorithms are used to get as accurate data and calculations as possible. In order to check the water quality, a reliable method had to be introduced. We have used python programming language to help calculate the quality of water using various machine learning algorithms that take the input of sensors set in place inside multiple overhead tanks. We use an ARIMA model to perform the above task. Auto-Regressive Integrated Moving Average (ARIMA) is a forecasting statistical technique that is used to analyze a time series based on history. A non-seasonal model is well-suited as it must be insensitive to any local short-lived trends within the time series. These trends do not contribute to the overall quality of water in the future.

In this manner we create a reliable and sustainable system that monitors the water quality to facilitate the enabling of a healthy and hygienic livelihood

We have used a Laptop with Microsoft Visual Studio Code application for this project to create a simple UI (User Interface) which has a number of buttons to perform the necessary calculations using various algorithms, Finally an graph comparing the success and errors of various algorithms are also defined to understand and select the best for determining water quality as accurately as possible.

A. Related Work

From the articles cited below we have learnt the importance of measuring the water quality to provide safe drinking water to the community including the various algorithms involved in doing so. The articles helped us understand the issues as well as merits in creating an system for forecasting and measuring the water quality.

The article at [1] indicated that due to increase population, advanced agricultural practices, industrialization, man- made activity, water is being highly polluted with different contaminants. Water is a vital resource for human survival. Also in [6] it is mentioned that the availability of good quality water is an indispensable feature for preventing diseases and improving quality of life. It is necessary to know details about different physico-chemical parameters such as colour, temperature, Total hardness, pH, sulphate, chloride, DO, BOD, COD, alkalinity used for testing of water quality.

The article at [2] elucidated on the fact that in recent days, the most important problem that our society faces is low quality of drinking water. Water quality monitoring is important because contaminated drinking water can spread diseases faster than any other sources. Articles [5,7,8] have expanded on the existing techniques, the general public is not aware of the potability of water. Lack of accurate and efficient low cost systems are a reason for poor awareness on the same. This paper focuses on modelling and developing a low cost water quality testing device and analysing its performance with the currently available products. The developed device can measure the parameters like pH, Total Dissolved Solids, Conductivity and Temperature. Its results are verified with samples of distilled water, salt water, tap water, dish wash and water, curd and performance is studied.

[3] has implemented a wireless sensor network for real time overhead tank water quality monitoring. Water is a precious source vital for healthy living. Most of the infectious diseases are due to contaminated water which leads to millions of deaths every year. There is a need to establish Water quality monitoring system to verify whether the determined water quality is suitable for intended use. This paper presents the application of Wireless Sensor Network (WSN) technology for real time online Water quality monitoring. In this paper, the details of system design and implementation of WSN are presented. Wireless Sensor Network (WSN) for a water quality monitoring is composed of number of sensor nodes with networking capability which are deployed at different overhead tanks and water bodies in an area. Each sensor node consists of an Arduino microcontroller, Xbee module and water quality sensors, the sensor probes shall continuously measure the different water quality parameters like pH, Temperature, Conductivity. The parameters are measured in real time by the sensors and send the data to the data center. Solar panel is used to power the system for each node. Data collected from remote nodes are displayed in the user PC. This developed system will demonstrate online sensor data analysis and has the advantages of power optimization, portability and easy installation. Similar experiments have been explored in [9,10]

[4] helped us understand that water is an essential natural resource that is fundamental to both human and animal health and more than 88% of South African households have access to water. Although South Africa has one of the cleanest water systems globally but there are still many factors causing water pollution.

II. METHODOLOGY

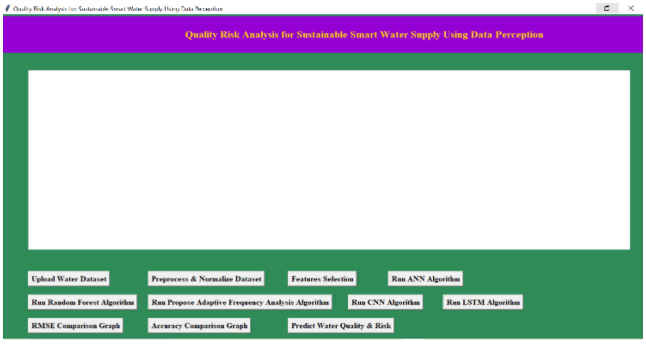

Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

Fig 2.1: Main Application Screen Layout

To get the above screen we have to click the run.bat file which is a Windows batch file. The basic layout of the screen is typed using Python code in Microsoft's Visual Studio Code application. It consists of a heading followed by a blank screen where the output will be shown along with various buttons which execute certain functions like uploading the dataset, various algorithms and comparison graphs.

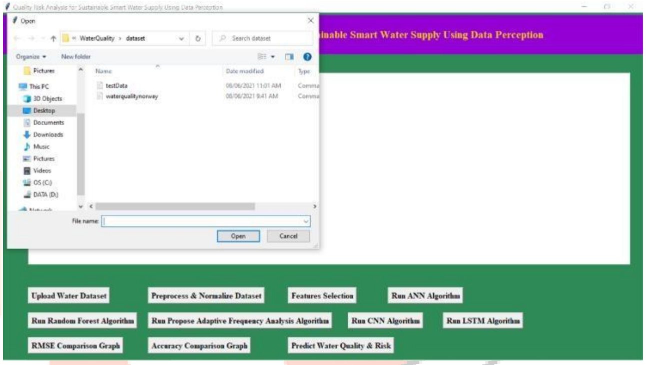
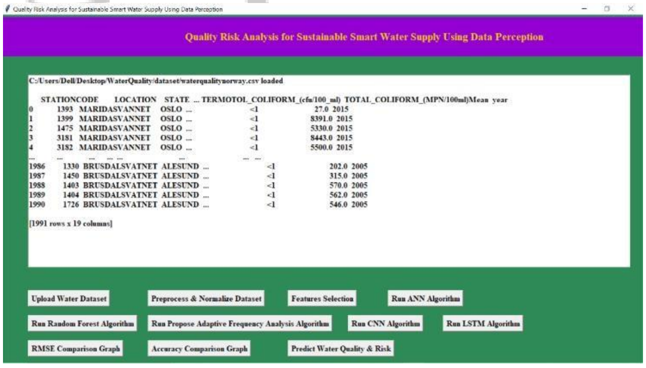


Fig 2.2: Selecting and Uploading the water dataset

The above water dataset containing some dummy values was taken from a reliable site called Kaggle which provides many such databases to work with for free. By clicking the "Upload water dataset", a pop up appears from which we have to select the required dataset.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

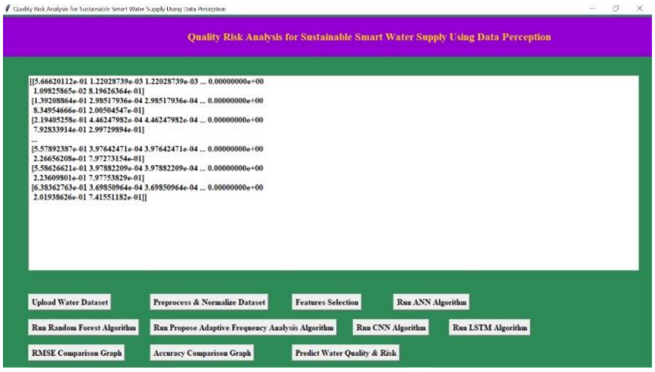
```
C:/Users/Dell/Desktop/WaterQuality/dataset/waterqualitynorway.csv loaded

   STATIONCODE      LOCATION   STATE ... TERMOTOL_COLIFORM_(cfu/100_ml) TOTAL_COLIFORM_(MPN/100ml)Mean  year
0         1393  MARIDASVANNET   OSLO ...                            <1                          27.0  2015
1         1399  MARIDASVANNET   OSLO ...                            <1                        8391.0  2015
2         1475  MARIDASVANNET   OSLO ...                            <1                        5330.0  2015
3         3181  MARIDASVANNET   OSLO ...                            <1                        8443.0  2015
4         3182  MARIDASVANNET   OSLO ...                            <1                        5500.0  2015
...        ...            ...    ... ...                           ...                           ...   ...
1986      1330 BRUSDALSVATNET ALESUND ...                           <1                         202.0  2005
1987      1450 BRUSDALSVATNET ALESUND ...                           <1                         315.0  2005
1988      1403 BRUSDALSVATNET ALESUND ...                           <1                         570.0  2005
1989      1404 BRUSDALSVATNET ALESUND ...                           <1                         562.0  2005
1990      1726 BRUSDALSVATNET ALESUND ...                           <1                         546.0  2005

[1991 rows x 19 columns]
```

Fig 2.3: Preprocess and Normalize the dataset to remove Null Values

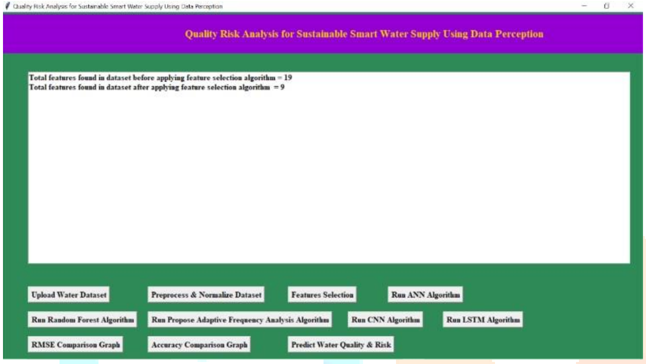
The above screen is shown once the dataset is selected. The white screen shows the names of rows and few of the data on columns along with the number of rows and columns in that particular dataset.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
[[5.66620112e-01 1.22028739e-03 1.22028739e-03 ... 0.00000000e+00
  1.09825865e-02 8.19626364e-01]
 [1.39208864e-01 2.98517936e-04 2.98517936e-04 ... 0.00000000e+00
  8.34954666e-01 2.00504547e-01]
 [2.19405258e-01 4.46247982e-04 4.46247982e-04 ... 0.00000000e+00
  7.92833914e-01 2.99729894e-01]
 ...
 [5.57892387e-01 3.97642471e-04 3.97642471e-04 ... 0.00000000e+00
  2.26656208e-01 7.97273154e-01]
 [5.58626621e-01 3.97882209e-04 3.97882209e-04 ... 0.00000000e+00
  2.23609801e-01 7.97753829e-01]
 [6.38362763e-01 3.69850964e-04 3.69850964e-04 ... 0.00000000e+00
  2.01938626e-01 7.41551182e-01]]
```

Fig 2.4: Feature Selection

To remove the Null values and non-numeric values that may exist in the dataset we have to click the "Preprocess and Normalize Dataset" button which converts Null/missing values(if any) and non-numeric values to numeric values. This helps the algorithms to run without any errors.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
Total features found in dataset before applying feature selection algorithm = 19
Total features found in dataset after applying feature selection algorithm  = 9
```

Fig 2.5: Resultant Data after Feature Selection

Once the Null values and non-numeric values are converted we have to click the "Features Selection" button to select only the important or relevant values from the dataset and remove the unimportant or irrelevant values. This helps to facilitate smooth operations of the various ML algorithms

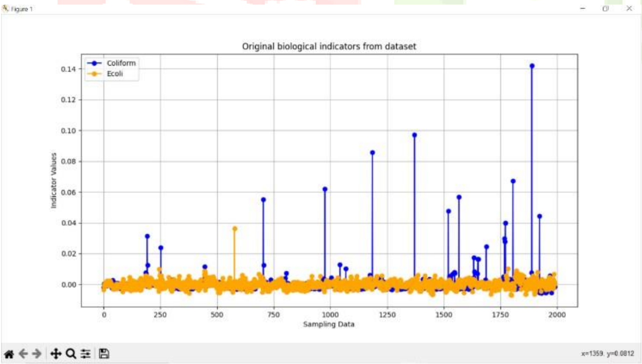
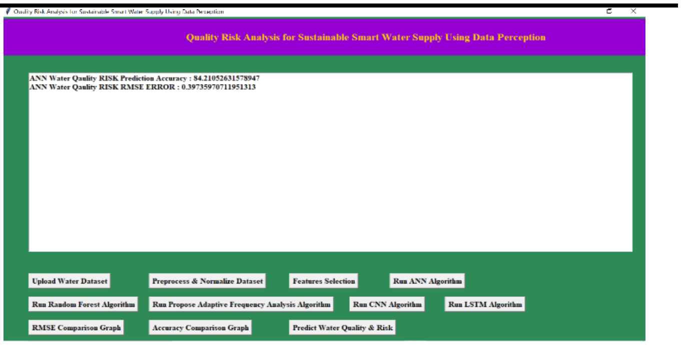


Fig 2.6: Graph Representing the various levels of bacteria present in water

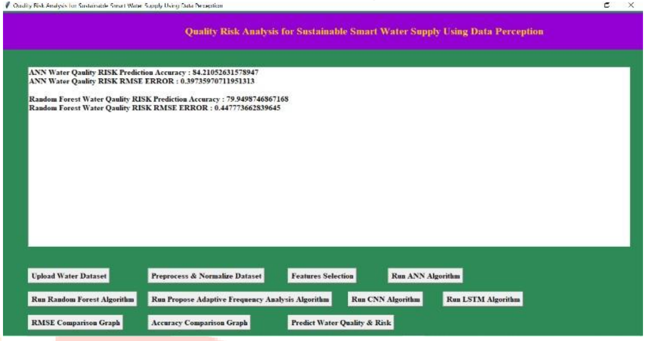
Once "Features Selection" is clicked a graph pops up which shows the level of 2 dangerous class of bacteria namely coliform and Ecoli which grow at room temperature and are hard to detect by conventional methods. Here Blue represents coliform bacteria and orange represents Ecoli.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
ANN Water Qaulity RISK Prediction Accuracy : 84.21052631578947
ANN Water Qaulity RISK RMSE ERROR : 0.39735970711951313
```

Fig 2.7: Resultant Data for ANN algorithm

After features selection we have to click the "Run ANN Algorithm" button to run the Artificial Neural Network (ANN) Algorithm on the data and we get the screen above showing the risk prediction accuracy as well as the root mean square error (RMSE) in calculating the risk.

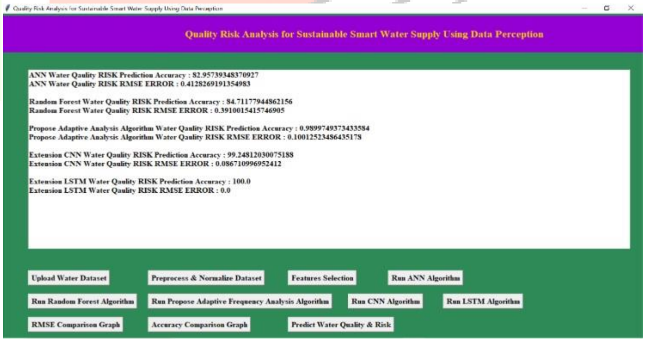

Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
ANN Water Qaulity RISK Prediction Accuracy : 84.21052631578947
ANN Water Qaulity RISK RMSE ERROR : 0.39735970711951313

Random Forest Water Qaulity RISK Prediction Accuracy : 79.9498746867168
Random Forest Water Qaulity RISK RMSE ERROR : 0.447773662839645
```

Fig 2.8: Resultant data for Random Forest Algorithm

After ANN Algorithm we have to click the "Run Random Forest Algorithm" button to run the Algorithm on the dataset and we get the screen above showing the risk prediction accuracy as well as the root mean square error (RMSE) in calculating the risk.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
ANN Water Qaulity RISK Prediction Accuracy : 82.95739348370927
ANN Water Qaulity RISK RMSE ERROR : 0.4128269191354983

Random Forest Water Qaulity RISK Prediction Accuracy : 84.71177944862156
Random Forest Water Qaulity RISK RMSE ERROR : 0.3910015415746905

Propose Adaptive Analysis Algorithm Water Qaulity RISK Prediction Accuracy : 0.9899749373433584
Propose Adaptive Analysis Algorithm Water Qaulity RISK RMSE ERROR : 0.1001252348643517S

Extension CNN Water Qaulity RISK Prediction Accuracy : 99.24812030075188
Extension CNN Water Qaulity RISK RMSE ERROR : 0.086710996952412

Extension LSTM Water Qaulity RISK Prediction Accuracy : 100.0
Extension LSTM Water Qaulity RISK RMSE ERROR : 0.0
```

Fig 2.9: Resultant data for Propose Adaptive Frequency Analysis and other algorithms

Similarly click other buttons to run the Convolutional Neural Network (CNN) algorithm, LSTM (Long-Short Term Memory) Algorithm and Adaptive Frequency Analysis and get the above screen showing the risk prediction accuracy and RMSE errors of each.

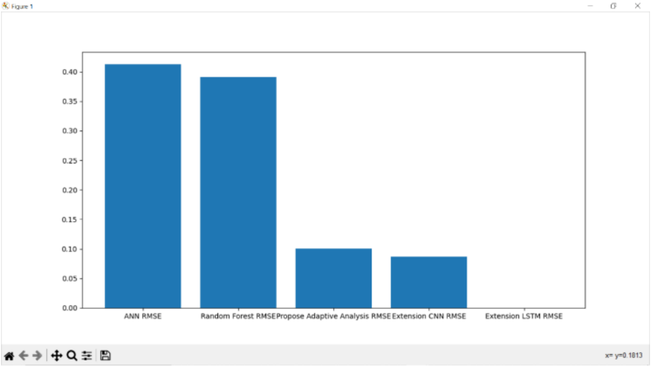


Fig 2.10: RMSE Comparison Graph

Once all the algorithms are executed on the dataset we get a screen containing respective values of accuracy and Root Mean Square Error (RMSE). Click the “RMSE Comparison graph” button to get the above graph which compares the RMSE errors of the various algorithms.

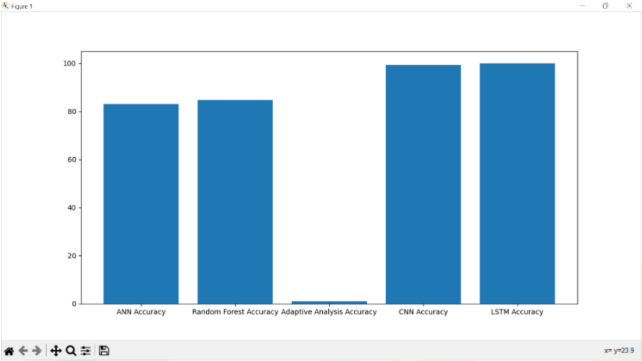


Fig 2.11: Accuracy Comparison Graph

Click the “Accuracy Comparison Graph” button to compare the risk prediction accuracies of various algorithms.

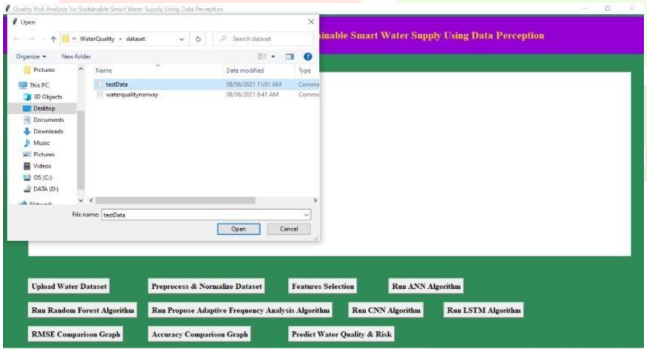
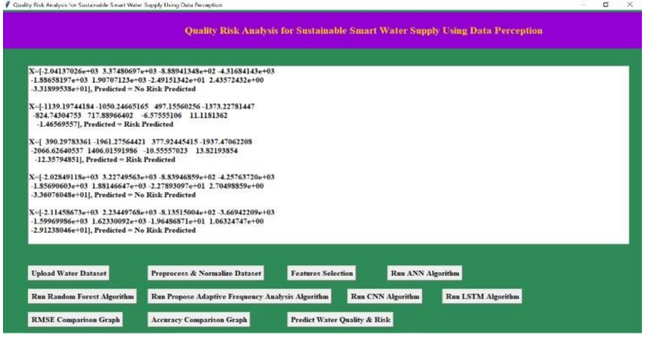


Fig 2.12: Uploading Test Data File

The ML algorithms that we run work on large amounts of data containing thousands of columns. Now we will try with smaller data containing arbitrary values set by us and upload that by clicking the “Predict Water Quality & Risk” button.


Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception
Upload Water Dataset
Preprocess & Normalize Dataset
Features Selection
Run ANN Algorithm
Run Random Forest Algorithm
Run Propose Adaptive Frequency Analysis Algorithm
Run CNN Algorithm
Run LSTM Algorithm
RMSE Comparison Graph
Accuracy Comparison Graph
Predict Water Quality & Risk

```
X=[-2.04137026e+03  3.37480697e+03 -8.88941348e+02 -4.31684143e+03
 -1.88658197e+03  1.90707123e+03 -2.49151342e+01  2.43572432e+00
 -3.31899538e+01], Predicted = No Risk Predicted

X=[-1139.19744184 -1050.24665165   497.15560256 -1373.22781447
  -824.74304753   717.88966402    -6.57555106    11.1181362
    -1.46569557], Predicted = Risk Predicted

X=[  390.29783361 -1961.27564421   377.92445415 -1937.47062208
 -2066.62640537  1406.01591986   -10.55557023    13.82193854
   -12.35794851], Predicted = Risk Predicted

X=[-2.02849118e+03  3.22749563e+03 -8.83946859e+02 -4.25763720e+03
 -1.85690603e+03  1.88146647e+03 -2.27893097e+01  2.70498859e+00
 -3.36076048e+01], Predicted = No Risk Predicted

X=[-2.11458673e+03  2.23449768e+03 -8.13515004e+02 -3.66942209e+03
 -1.59969986e+03  1.62330092e+03 -1.96486871e+01  1.06324747e+00
 -2.91238046e+01], Predicted = No Risk Predicted
```

Fig 2.13: Predicted Water Quality and Risk on test data

After test data is uploaded it manually calculates it for each and every column value. Hence why this is used for smaller test files as the larger the file, the calculations become more complex.

III. RESULTS AND DISCUSSIONS

This study reflects on the implementation of a water quality monitoring and forecasting system based on machine learning and artificial intelligence where we use multiple algorithms to test the water quality and predict the risk associated with it as accurately as possible.

The results are shown in the figures 2.9, 2.10, 2.11 and 2.13.

IV. CONCLUSIONS

We know that water is a very essential resource and is vital for the survival of the human civilization along with other living beings. We are living in a world that rapidly advances in technology day by day and the main aim is to provide safe drinkable and useable water to the living beings. To do so various industries are aiming to provide safe drinking water to various communities. Through the use of this algorithm we have seen the monitoring and forecasting of water quality based on data given/measured by us.

References

[1] Patil, P. N., D. V. Sawant, and R. N. Deshmukh. "Physico- chemical parameters for testing of waterA review." International Journal of Environmental Sciences 3.3 (2012): 1194.

[2] Indu, K., and Jishmi Jos Choondal. "Modeling, development & analysis of low cost device for water quality testing." 2016 IEEE Annual India Conference (INDICON). IEEE,2016.

[3] Sowmya, Ch, et al. "Implementation of wireless sensor network for real time overhead tank water quality monitoring." 2017 IEEE 7th International Advance Computing Conference (IACC). IEEE, 2017.

[4] Rasin, Zulhani, and Mohd Rizal Abdullah. "Water quality monitoring system using zigbee based wireless sensor network." International Journal of Engineering &Technology 9.10 (2009): 24-28

[5] Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time Series Analysis: Forecasting and Control;John Wiley & Sons: Hoboken, NJ, USA, 2015.

[6] Tyralis, H.; Papacharalampous, G. Variable selection in time series forecasting using random forest algorithms 2017,10, 114.

[7] Hernández, Nathalie, et al. "Arima as a forecasting tool for water quality time series measured with UV-Vis spectrometers in a constructed wetland." Tecnología y ciencias del agua 8.5 (2017): 127-139.

[8] An, Qi, and Min Zhao. "Time Series Analysis in the Prediction of Water Quality." 7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017). Atlantis Press, 2017.

[9] K. H. Kamaludin and W. Ismail, "Water quality monitoring with internet of things (IoT)," 2017 IEEE Conference on Systems, Process and Control (ICSPC), Malacca, 2017,pp. 18-23.

[10] N. Kumar Koditala and P. Shekar Pandey, "Water Quality Monitoring System Using IoT and Machine Learning," 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, 2018, pp. 1-5

[11] Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text \classifiers. In SIGIR'94 (pp. 3-12). Springer, London.

[12] Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval(Vol. 33, pp. 81-93).

[13] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419-444.

[14] Mahinovs, A., Tiwari, A., Roy, R., & Baxter, D. (2007). Text classification method review.

[15] Mund, S. (2015). Microsoft azure machine learning. Packt Publishing Ltd.

[16] Rogati, M., & Yang, Y. (2002, November). High-performing feature selection for text classification. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 659-661). ACM.

[17] Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering.In 2005 International Conference on Cyberworlds (CW'05) (pp. 4- pp). IEEE. 50 65.

[18] Sasaki, M., & Shinnou, H. (2005, November). Spam detection using text clustering. In null (pp. 316-319) IEEE. 26) Sebastiani, F. (2005). Text categorization. In Encyclopedia of Database Technologies and Applications (pp. 683-687). IGI Global.

[19] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[20] Sorkin, D., E. (n.d.). Learn how to identify spam.

[21] Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), 45-66.

[22] Yerazunis, W. S. (2004, January). The spam-filtering accuracy plateau at 99.9% accuracy and how to get past it. In Proceedings of the 2004 MIT Spam Conference