International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 15th Dec 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 12

DOI: 10.48047/IJIEMR/V11/ISSUE 12/24

Title **DECISION TREE AND SUPPORT VECTOR MACHINE FOR ANOMALY DETECTION IN WATER DISTRIBUTION NETWORKS**

Volume 11, ISSUE 12, Pages: 169-179

Paper Authors
Narayanadasu Radhika, Dr. G Venkata Rami Reddy

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

DECISION TREE AND SUPPORT VECTOR MACHINE FOR ANOMALY DETECTION IN WATER DISTRIBUTION NETWORKS

1 Narayanadasu Radhika, M.tech in COMPUTER SCIENCE (CS) SIT JNTUH
2 Dr. G Venkata Rami Reddy, Professor of IT

ABSTRACT: Monitoring the quality of the drinking water is crucial nowadays since the water supply may be contaminated and can spread a number of diseases. Therefore, it is essential to promptly identify contamination and prevent any entry into water distribution systems. We have a variety of machine learning methods for classification to address issues with intrusion detection, but picking the optimal one is a crucial effort. We did a trial research on machine learning methods to decide the ideal calculation for our water quality checking framework. In this trial work, we utilized a genuine dataset got from a Tunisian water treatment office to look at the presentation of the two notable order strategies in the literature, Decision Tree and Support Vector Machines.

Keywords *– Support Vector Machine, Decision tree and Recurrent neural networks.*

1. INTRODUCTION

Water utilized for human utilization should be liberated from unsafe microbes or perilous substances. Consequently, it is fundamental for stop any passage into water circulation frameworks and to recognize any deliberate or unexpected contamination when attainable. Conventional water quality observing involves physically gathering tests from different water conveyance network areas, which are thusly conveyed to assigned research centers for defilement testing [1, 2]. To quantify the different water quality markers, the water treatment station "Ghédir El Golla" in Tunisia utilizes free, compact recognition tests that should be lowered in water sources. In small towns and cities, physical, chemical, and microbiological testing are performed once a week and twice a week, respectively. However, because it is costly, time-consuming, and lacks immediate feedback, the conventional method to water quality regulation is incredibly ineffective. The terrible effects of water contamination [3] need a more expedient and affordable solution. Today, one of the biggest issues facing smart cities is the smart and ongoing monitoring of water quality. Using wireless sensor technology, the quality of drinking water may be continuously monitored. Applications that use wireless sensor networks (WSNs) must continually adapt to changes in the environment, hardware deterioration, and faulty sensor readings.

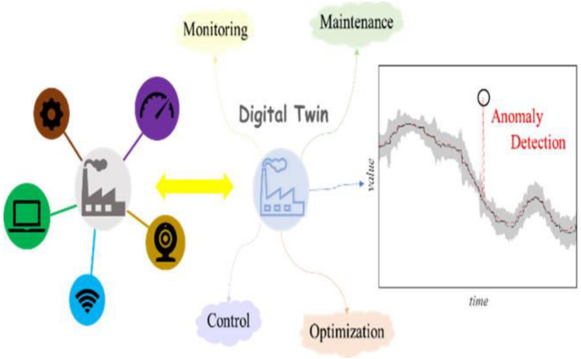


Fig.1: Example figure

Consequently, a WSN program much of the time needs to find out about and conform to changes in its working climate to hold fitting functional exactness. These issues have been tended to with the utilization of machine learning (ML). The creation and improvement of calculations that empower frameworks to utilize exact information, experience, and preparing to develop and adjust to changes that happen in their current circumstance is a vital accentuation of ML research. Various machine learning (ML) procedures have been utilized in an assortment of sensor network applications, like observing water quality.

2. LITERATURE REVIEW

2.1 Monitoring source water for microbial contamination: evaluation of water quality measures

Microbial water quality has an effect on human, animal, and environmental health. Multidisciplinary research focuses on microbiological water quality monitoring, prediction, and management. This special collection of papers in the was inspired by the idea of creating a separate section that would cover the problems and advances in microbial water quality research from a broader perspective. It discusses a variety of aspects of the release, movement, and environmental survival of microorganisms that are associated with human health. The papers look at the spatiotemporal variety of microbial water quality, the selection of indicators of the spatiotemporal varieties, the capability of base residue and biofilms, relationships between's groupings of pointer and pathogenic life forms and the capability of chance evaluation strategies, utilization of sub-atomic markers, subsurface microbial vehicle as connected with microbial water quality, anti-toxin obstruction, constant observing and nowcasting, and watershed scale model. International expertise in the topic is represented by both the writers and the editors. The results highlight the difficulties in monitoring and comprehending microbiological water quality; they also offer interesting study possibilities for enhancing the body of information required to safeguard and enhance our water supplies.

2.2 Contamination Potentials of Household Water Handling and Storage Practices in Kirundo Subcounty, Kisoro District, Uganda

In emerging and underdeveloped nations, waterborne illnesses are a significant public health burden. Public health is at danger when polluted water is consumed, and the problem is particularly concerning in rural regions. In the Kirundo subcounty, Kisoro Locale, Uganda, the objective of this study was to assess the defilement possibilities of different family water dealing with and capacity procedures. Materials and strategies: In a cross-sectional and clear review, 344 water tests were haphazardly chosen, dissected utilizing the Most Probable Number(MPN) strategy for bacterial defilement, all total coliforms (TCs), and Escherichia coli per 100 ml, and the outcomes were accounted for with regards to CFU/100 ml. Results: Escherichia coli was available in 34.1% of the examples from unprotected water sources and all out coliforms in 43.2% of them. 25% of the home drinking water tested positive for total coliforms, while 8.7% tested positive for Escherichia coli. It was discovered that the majority of drinking water sources had coliform levels higher than what was advised by national and international norms. Regarding total coliforms and Escherichia coli, there was statistically significant variation across the water sources (p 0.05). The safeguarded water source was more secure than unprotected water sources, as per the general discoveries, which showed a significant relationship between's microbiological water quality and water source neatness. For the community's vulnerable water sources, monitoring and protection procedures are advised.

2.3 Surface Water Pollution Detection using Internet of Things

One of the basic necessities and essential to maintaining quality of life is water. Due to

Pakistan's rural economy, its relevance is more than usual. The development of urbanisation and industrialisation is causing a constant decline in water quality. To achieve this, we suggest an Internet of Things (IoT)-based water quality system that can measure water quality almost instantly. The suggested remedy is based on water quality parameters set out by the World Health Organization (WHO). A real-time embedded prototype has been created specifically for this purpose and will record the water quality characteristics from the water samples that have been gathered from various sources around the research area. The hardware solution transmits data to the cloud for instant processing and archiving. Our created software solution, which consists of a mobile app and a dashboard, allows for remote monitoring and control of the processed data as well as the regulation of water flow. Additionally to a system for monitoring and controlling water quality, a predictive analysis of the gathered data has been carried out. A dataset has been received from Pakistan Council of Research on Water Resources for training purposes (PCRWR). Deep neural networks surpass all other techniques with an accuracy of 93%, according to experimental data using machine learning algorithms for classifying water quality. The early findings indicate that there is great potential for building up this idea to an advanced level.

2.4 Real-time Learning-based Monitoring System for Water Contamination

Cities need real-time monitoring of water quality because it has a big effect on people's health. Unfortunately, it is not possible for humans to regularly, much alone in real time, examine the chemical composition of water due to how difficult this would be to do. So, using a system that is really effective can be a great answer. The framework we created in this work, the Ongoing Wise Checking Framework for Water Quality, empowers a city to answer potential flare-ups of contamination and to protect its residents. The system has the capacity to process and categorise the information retrieved from visual pictures, which will significantly save costs and manpower. In this instance, two features are presented for saliency features and colour features, respectively: Fast Fourier Transform (FFT) and Color Layout Descriptor (CLD). While CLD is able to represent the colour data with great efficacy and efficiency, FFT performs well at extracting saliency information and is not computationally costly. Additionally, this method makes use of Support Vector Machine (SVM), which has the advantages of training quickly, using little data, and accurately classifying instances of floating trash and other forms of water pollution. The accuracy has achieved 75% up until this point, which is encouraging. The effective features & classifiers would serve as strong approaches to automatically monitor water contamination, even though the detection performance can still be enhanced.

3. METHODOLOGY

An IoT-based answer for track the water quality is presented in the current framework. The recommended framework utilizes a portable application to convey remote observing of water quality assessment and water stream the executives. For the arrangement of water quality, four AI calculations — upport Vector Machine (SVM), K-Nearest Neighbor (KNN), single-layer neural network, and deep neural network — have been utilized. Just three factors — turbidity, temperature, and pH — that are estimated as per What criteria's identity is utilized to decide the water quality in this review.

Disadvantages:

Below, we When estimating water quality, using only three factors and comparing them to established values is quite a restriction.

We have a variety of machine learning algorithms for classification in our project's intrusion detection proposal, but picking the optimal one is a crucial issue. We did a trial research on machine learning strategies to decide the ideal calculation for our water quality checking framework. In this exploratory work, we utilized a genuine dataset got from a Tunisian water treatment office to look at the exhibition of the two notable grouping strategies in the writing, Decision Tree and Support Vector Machines.

Advantages:

We compared the Decision Tree and Support Vector Machine classification techniques that are well-known in the literature in terms of accuracy, precision, and recall.

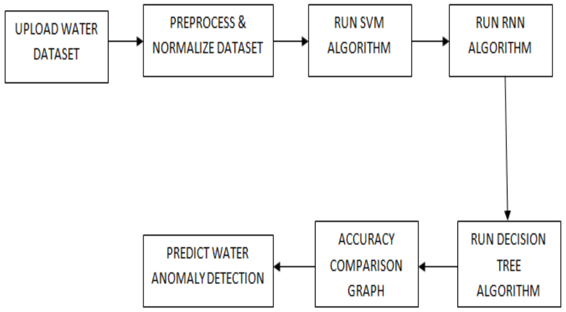

UPLOAD WATER DATASET
PREPROCESS & NORMALIZE DATASET
RUN SVM ALGORITHM
RUN RNN ALGORITHM
RUN DECISION TREE ALGORITHM
ACCURACY COMPARISON GRAPH
PREDICT WATER ANOMALY DETECTION

Fig.2: System architecture

MODULES:

In order to carry out this project, we created the following modules.

1) **Upload Water Dataset:** This module will be used to upload a water dataset to the application.
2) **Preprocess & Normalize Dataset:** Since the dataset frequently contains missing and non-numeric values, we will use the preprocessing approach to replace them with integer ids before using MEAN to normalise all of the values. Following preprocessing, the dataset will be parted into train and test segments, with 80% of the dataset being utilized to prepare AI calculations and 20% being utilized to test and estimate the precision of the prepared strategy.
3) **Run SVM Algorithm:** With the use of the given dataset and this module, we will train the SVM algorithm and create an SVM model. On test data, this model will be used to calculate SVM accuracy, precision, and recall.
4) **Run RNN Algorithm:** With the help of this module, we will train the RNN algorithm with the aforementioned dataset before creating the RNN model. This model will be used to calculate the RNN's accuracy, precision, and recall using test data.
5) **Run Decision Tree Algorithm:** With the use of the given dataset, we will train the decision tree method in this module, which will subsequently be used to create the decision tree model. This model will be used to calculate Decision Tree accuracy, precision, and recall using test data.
6) **Accuracy Comparison Graph:** This module will be used to create a comparison graph between all three methods.
7) **Predict Water Anomaly Detection:** This module will allow us to submit test data, and machine learning will

subsequently determine whether or not an anomaly is there.

4. IMPLEMENTATION

In this article, the author uses a variety of machine learning methods, such as SVM and Decision Tree, to forecast water pollution or the presence of anomalous ECOLI bacteria in water. However, the student requested to use RNN, so we utilised RNN subversion methodology called LSTM (Long Short Term Memory).

We utilised the author's dataset to train the aforementioned method.

ALGORITHM:

SVM:

One of the most popular regulated learning calculations, Support Vector Machine, or SVM, is utilized to address Characterization and Relapse issues. Nonetheless, it is generally utilized in Machine Learning Order issues. The SVM calculation's goal is to lay out the ideal line or choice limit that can separate n-layered space into classes, permitting us to rapidly arrange new data of interest from here on out. A hyperplane is the name given to this ideal choice limit. SVM chooses the outrageous vectors and focuses that guide in the making of the hyperplane. Support vectors, which are utilized to address these outrageous occasions, structure the reason for the SVM strategy. Consider the image underneath, where a choice limit or hyperplane is utilized to sort two unmistakable classes.

RNN:

In voice acknowledgment and normal language handling, rrecurrent neural networks (RNNs) are a type of artificial neural network (NLP). Deep learning and the production of models that imitate the terminating of neurons in the human mind both utilize RNN. Repetitive organizations are made to distinguish designs in information arrangements, including text, genomes, penmanship, communicated in language, and mathematical time series information from sensors, stock trades, and administrative associations. A memory-state is added to the neurons in a recurrent neural network, which otherwise resembles a conventional neural network. A basic memory is to be used in the computation. A deep learning-focused technique that employs a sequential method is the recurrent neural network. We always assume in neural networks that every input and every output is dependent on every other layer. Recurrent neural networks are so named because they carry out mathematical operations in a consecutive manner.

DECISION TREE:

A supervised learning strategy called a decision tree is utilized in information digging for techniques for characterization and relapse. We can utilize this tree to support direction. The characterization or relapse models are created as a tree structure by the decision tree. It partitions an informational index into more modest subgroups while gradually fabricating the choice tree. The decision nodes and leaf hubs make up the last tree. There are no less than two branches on a decision node. The classification or decision is shown in the leaf nodes. We are restricted in our capacity to partition leaf hubs further. The root node is the highest decision node in a tree and is associated with the best indicator. All out and mathematical information may both be taken care of by decision trees.

5. EXPERIMENTAL RESULTS

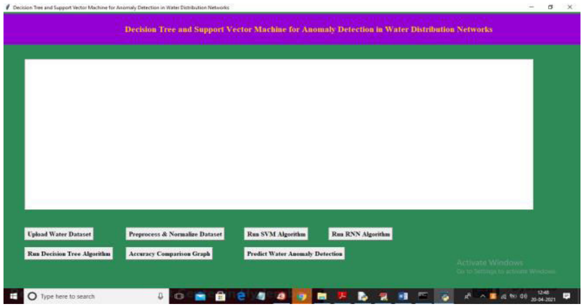
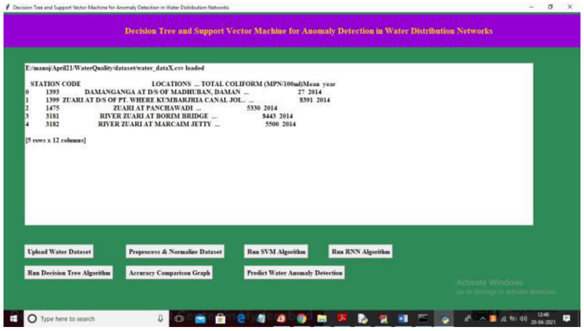

Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Type here to search

Fig.3: Home screen

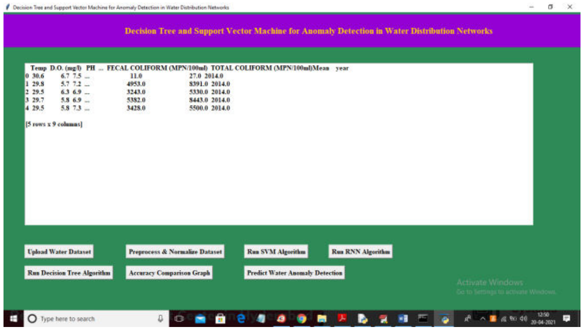

Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Type here to search

```
E:/manoj/April21/WaterQuality/dataset/water_dataX.csv loaded

  STATION CODE                    LOCATIONS ... TOTAL COLIFORM (MPN/100ml)Mean  year
0      1393        DAMANGANGA AT D/S OF MADHUBAN, DAMAN ...            27  2014
1      1399  ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI... ...     8391  2014
2      1475                  ZUARI AT PANCHAWADI ...        5330  2014
3      3181              RIVER ZUARI AT BORIM BRIDGE ...        8443  2014
4      3182              RIVER ZUARI AT MARCAIM JETTY ...        5500  2014

[5 rows x 12 columns]
```

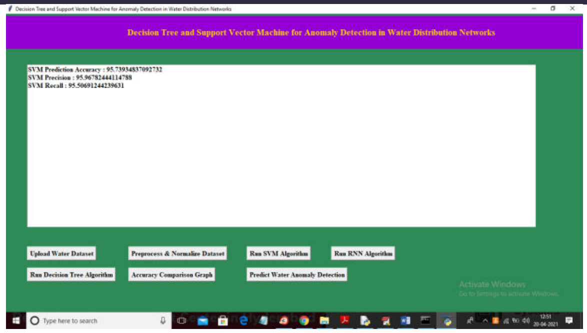
Fig.4: Upload water dataset


Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Type here to search

```
  Temp  D.O. (mg/l)  PH ... FECAL COLIFORM (MPN/100ml)  TOTAL COLIFORM (MPN/100ml)Mean   year
0  30.6       6.7  7.5 ...           11.0              27.0  2014.0
1  29.8       5.7  7.2 ...         4953.0            8391.0  2014.0
2  29.5       6.3  6.9 ...         3243.0            5330.0  2014.0
3  29.7       5.8  6.9 ...         5382.0            8443.0  2014.0
4  29.5       5.8  7.3 ...         3428.0            5500.0  2014.0

[5 rows x 9 columns]
```

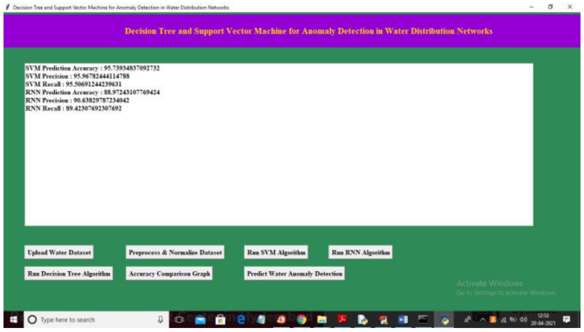
Fig.5: Preprocess & normalize dataset



Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:51
20-04-2021

```
SVM Prediction Accuracy : 95.73934837092732
SVM Precision : 95.96782444114788
SVM Recall : 95.50691244239631
```

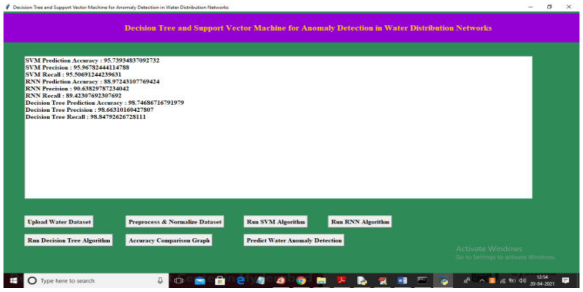
Fig.6: SVM algorithm



Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:53
20-04-2021

```
SVM Prediction Accuracy : 95.73934837092732
SVM Precision : 95.96782444114788
SVM Recall : 95.50691244239631
RNN Prediction Accuracy : 88.97243107769424
RNN Precision : 90.63829787234042
RNN Recall : 89.42307692307692
```

Fig.7: RNN algorithm



Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:54
20-04-2021

```
SVM Prediction Accuracy : 95.73934837092732
SVM Precision : 95.96782444114788
SVM Recall : 95.50691244239631
RNN Prediction Accuracy : 88.97243107769424
RNN Precision : 90.63829787234042
RNN Recall : 89.42307692307692
Decision Tree Prediction Accuracy : 98.74686716791979
Decision Tree Precision : 98.66310160427807
Decision Tree Recall : 98.84792626728111
```

Fig.8: Decision tree algorithm

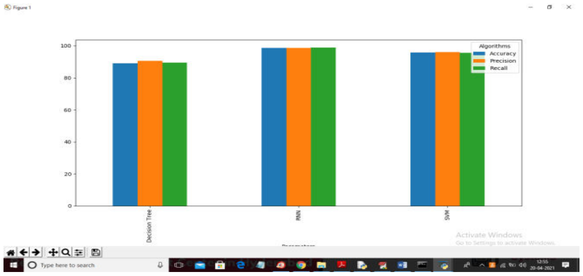
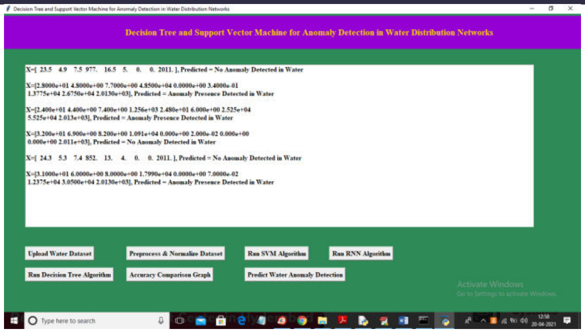


Fig.9: Accuracy comparison graph



Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks
Upload Water Dataset
Preprocess & Normalize Dataset
Run SVM Algorithm
Run RNN Algorithm
Run Decision Tree Algorithm
Accuracy Comparison Graph
Predict Water Anomaly Detection
Activate Windows
Go to Settings to activate Windows.
Type here to search
12:58
20-04-2021

```
X=[ 23.5   4.9   7.5 977.   16.5   5.    0.    0.  2011. ], Predicted = No Anomaly Detected in Water

X=[2.8000e+01 4.8000e+00 7.7000e+00 4.8500e+04 0.0000e+00 3.4000e-01
 1.3775e+04 2.6750e+04 2.0130e+03], Predicted = Anomaly Presence Detected in Water

X=[2.400e+01 4.400e+00 7.400e+00 1.256e+03 2.480e+01 6.000e+00 2.525e+04
 5.525e+04 2.013e+03], Predicted = Anomaly Presence Detected in Water

X=[3.200e+01 6.900e+00 8.200e+00 1.091e+04 0.000e+00 2.000e-02 0.000e+00
 0.000e+00 2.011e+03], Predicted = No Anomaly Detected in Water

X=[ 24.3   5.3   7.4 852.   13.    4.    0.    0.  2011. ], Predicted = No Anomaly Detected in Water

X=[3.1000e+01 6.0000e+00 8.0000e+00 1.7990e+04 0.0000e+00 7.0000e-02
 1.2375e+04 3.0500e+04 2.0130e+03], Predicted = Anomaly Presence Detected in Water
```

Fig.10: Predict water anomaly detection

6. CONCLUSION

Using a real dataset, this research suggested an investigation of the effectiveness of two well-known classification methods in the literature: Decision Trees and SVMs. The test results provided us with solid evidence of the effectiveness of the categorization approaches. Additionally, we discovered that linear SVM appears to be suitable for our system of water quality monitoring. We are attempting to incorporate a novel data aggregation method in order to reduce the amount of data required to run the SVM classification method for future research.

REFERENCES

[1] DP. SARTORY and J. WATKINS, “Conventional culture for water quality assessment: is there a future?,” Journal of applied microbiology, 1998, vol. 85, no S1, pp. 225S-233S.

[2] JD. PLUMMER and SC. LONG, "Monitoring source water for microbial contamination: evaluation of water quality measures," Water research, 2007, vol. 41, no 16, pp. 3716-3728.

[3] Agensi, Alexander, et al., "Contamination Potentials of Household Water Handling and Storage Practices in Kirundo Subcounty, Kisoro District, Uganda," Journal of environmental and public health, 2019, vol. 2019.

[4] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar and H. Khurshid, "Surface Water Pollution Detection using Internet of Things," 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, 2018, pp. 92-96.

[5] Q. Chen et al., "Real-time Learning-based Monitoring System for Water Contamination," 2018 4th International Conference on Universal Village (UV), Boston, MA, USA, 2018, pp. 1-5.

[6] V. A. Usachev, L. I. Voronova, V. I. Voronov, I. A. Zharov and V. G. Strelnikov, "Neural Network Using to Analyze the Results of Environmental Monitoring of Water," 2019 Systems of Signals Generating and Processing in the Field of on Board Communications, Moscow, Russia, 2019, pp. 1-6.

[7] République Tunisienne Ministère de L'agriculture, Société Nationale d'exploitation et de Distribution des eaux. Projet d'approvisionnement en eau potable du grand Tunis et des centres urbains et financement additionnel. Ref : 289/14. Version finale Janvier 2015.

[8] NB. AMOR, S. BENFERHAT and Z. ELOUEDI, "Naive bayes vs decision trees in intrusion detection systems," Proceedings of the 2004 ACM symposium on Applied computing. ACM, 2004. pp. 420- 424.

[9] T. AMBWANI, "Multi class support vector machine implementation to intrusion detection," Proceedings of the International Joint Conference on Neural Networks, 2003. IEEE, 2003. pp. 2300-2305.

[10] Hao, Qi, and Fei Hu, "Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing, and Machine Learning," Taylor & Francis, 2012.

[11] V. N. VAPNIK., "Statistical Learning Theory," New York, wiley édition, 1998.

[12] BE. BOSER, IM. GUYON and VN. VAPNIK, “A training algorithm for optimal margin classifiers,” Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992. pp. 144- 152.

[13] VN. VAPNIK, “The nature of statistical learning theory,” SpringerVerlag, 1995.

[14] J. PLATT, “Fast Training of Support VectorMachines using SequentialMinimal Optimization, Advances in Kernel MethodsSupport Vector Learning,” MIT Press, 1999, pp. 185–208.

[15] T. JOACHIMS, “Estimating the Generalization Performance of a SVM Efficiently,” ICML-00, 2000, pp. 431–438.