Contamination Detection In Water Distribution

***Abstract*—Water resources are an indispensable and important resource in human daily life, and the water distribution system can provide human beings with a large amount of high-quality water resources. However, how to ensure the quality of drinking water in the water distribution system is a major issue that society attaches great importance to. The water distribution system continuously generates a large amount of water quality data, such as water temperature, ph value, etc., which are usually only about one minute apart. Manual monitoring of water quality data is not feasible. Many water distribution companies are working on using AI technology to automatically detect abnormal water quality data. . But the problem also comes from this. Although AI technology reduces a lot of costs compared with manual monitoring methods, whether the accuracy of detecting abnormal data is higher than that of manual methods is something that water companies need to consider.**

This article will use a variety of machine learning algorithms (SVM, ANN, LSTM), compared and evaluated high-quality real-time data analysis methods that can detect abnormal water quality data in real time. To ensure the feasibility of the experiment, the data will be selected with time attributes and extracted from the real world. Because the data is highly unbalanced, accuracy is not a comparison criterion that can be used in this article. The F1 score will be used in this article to discuss whether algorithms can be integrated into detection systems and alert when water quality data deviates from normal.

Keywords- water distribution system, water quality features, machine learning, anomaly detection, F1 score.

I. Introduction

Human life cannot be separated from water. A person can go without food for three days, but if a person is asked not to drink water for one day, it is very likely to cause death. And high-quality water is of great benefit to people and society. For people, high-quality water resources can ensure people's health and prolong people's life expectancy. For society, it can reduce the burden on regional health care facilities and improve the national economy. But in daily life, how can people get high-quality drinking water? This has to mention the water distribution system.

Water distribution systems are a major asset for water utilities. AWWA(American Water Works Association) defines it as a water utility component that distributes finished drinking water to customers [1,]. But most water distribution systems only meet the treatment plant's drinking water standards, ignoring the possible deterioration of water quality during distribution. Water-borne diseases remain a major health concern. According to the WHO 1999 Burden of Disease Measure Report, 2.4 million people suffer from diarrhoeal disease and some die each year due to contaminated drinking water [2,]. This health burden is largely borne by populations and children in developing countries.

There are many reasons for the pollution of water quality in the water distribution system. The change of hydraulic parameters may be due to the leakage of the system, which is likely to cause external microorganisms to enter the system [3,]. Most of the water-borne diseases mentioned above are caused by system leakage. of. In addition to microbial issues in drinking water, water security is also affected by many factors, including chemistry. And these problems are usually not directly noticed by the drinker. The intermediary water companies, while gaining the trust of their customers, should be clear about their responsibilities: ensuring the health of those who use the water or services they provide.

The basis for the water distribution company to judge whether the water quality is correct is whether all its values are within the normal range, and there are many ways to detect water pollution. The most widely used method today is laboratory analysis, in which water company staff collect water samples at specified times and send them to the laboratory for analysis. This method has very high accuracy under the work of highly skilled personnel, but also has considerable limitations, such as high labor cost and very time-consuming. Compared with the traditional manual method, the machine learning method has been widely discussed whether it can be used in the water distribution system to detect abnormal values since it was proposed. The GECCO competition posed a related question in 2018, whether it is possible to use machine learning algorithms to analyze water quality data [4,]. The research question of this paper also comes from this, how to develop a real-time data analysis method to detect water pollution problems. After doing a lot of research, I will use SVM(Support Vector Machine), ANN(Artificial Neural Network) in traditional machine learning algorithms and LSTM(Long-Short-Term Memory) in deep learning algorithms for comparative evaluation, so as to obtain an effective data analysis method.

II. Background

A. Anomaly Detection

In order to extract anomalous data from a dataset, anomaly detection methods are required. Research on anomaly detection has lasted for decades, and many important detection technologies have been developed, among which the application of machine learning technology has greatly improved the ability of anomaly detection. Research on anomaly detection problems involves engineering, statistics, machine learning, data mining and other disciplines [1,]

Anomaly detection refers to "the problem of finding patterns in data that do not conform to expected behavior" [2,]. If we set χ be the data space for some specific application, then the anomalies can be defined as:

$$A = \{x \in \chi | p^{+}(x) < \tau\}, \tau \geq 0$$

τ is the threshold such we will define the x as abnormal.
For different applications, different techniques have been proposed to deal with anomalous data, and after decades of research, a variety of detection techniques have been proposed. [3,] Failure to employ anomaly detection to protect data can lead to very serious and uncontrollable data risk. One example is that when the server is receiving the data sent by the client, if the data is not detected, the hacker can send virus data to the server and cause the server to crash [4,]; another example is when detecting the flight attitude of the aircraft, if Without data detection, abnormal data may lead to the crash of the aircraft.
Although anomaly detection is very important to our lives, it is challenging to identify ano malous data from normal data. On the one hand, due to the high variability of normal data, we may misclassify normal data as abnormal data, and we may also misclassify abnormal data as normal data. For example, variation in biological data often occurs, so we are likely to classify the mutated normal data as abnormal data. In addition, the noise contained in the dataset may also mask the abnormal characteristics of the abnormal data. On the other hand, the quality of the training dataset will also affect the results of anomaly detection, because the abnormal data is after all a minority in the entire dataset, so the training set itself is likely to contain very few anomaly datasets, and the final training results will be very bad for the recognition of anomalous data. [1,] There are three categories of anomaly detection, there are point anomalies, context c, and collective anomalies, each of which is defined as follows:

1) Point anomalies: A point anomaly is the simplest type of anomaly, which indicates that one piece of data is abnormal compared to other data.
2) Context anomalies: Context exception is that a data is abnormal in one context and normal in another context.
3) Collective anomalies: For the entire data set, if a subset of the data is abnormal under the entire data set, it is called a collective exception.

By comprehensively reading the previous literature on anomaly detection, I found that when using machine learning to deal with anomaly detection problems, SVM is the most commonly used model [1,]. Machine Learning and Deep Learning has play an important role in anomaly detection.

B. Machine Learning

Machine learning is now closely related to our lives, in our daily life can see the application of machine learning everywhere. When we use mobile phones for voice input, automatic navigation or online shopping, machine learning will come into play. Machine learning makes our lives more convenient, secure and efficient. There are two main types of machine learning algorithm, the supervised machine learning and the unsupervised machine learning, they are define as follows [5,]:

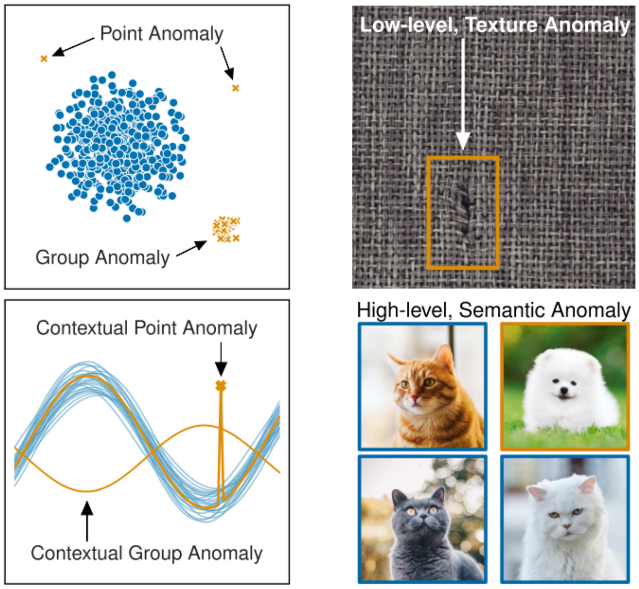


Fig. 1. Different types of anomalies.

- supervised machine learning are the most commonly used. The scientist must try to teach the algorithm to make correct decision with training data with labels. For example, we can teach the algorithm to distinguish different fruit by feeding the algorithm of pictures about different fruits with label on each picture. Some common supervised machine learning algorithms include linear and logistic regression, multiclass classification, and **support vector machines(SVM)**.
- unsupervised machine learning teaches the algorithm to recognize complex processes and patterns without providing a guidence. For example, if we use unsupervised machine learning to distinguish difference fruit, we would let the algorithm to find similarity and difference of each picture like colors and patterns. Then the algorithm can seperate the pictures into groups by their colors and patterns.

SVM: SVM is short for Support Vector Machine, it is a supervised learning model that analyzes data in classification and regression analysis and related learning algorithms. To put it simply, SVM is a two-class classification model. Its basic model is a linear classifier with the largest margin defined in the feature space. The learning strategy of SVM is to maximize the margin. In Fig. 2, we want to classify the black and white points in dimension 2. And there are three straight lines in the picture, their are used as a classifier, the question is which classifier is better? First, we can exclude H1 because it cannot separate the black points from the white points. Secondly, although H2 can successfully distinguish all black pointss from white pointss, we think it is not as good as H3, because the distance between H2 and the nearest point is very close. As

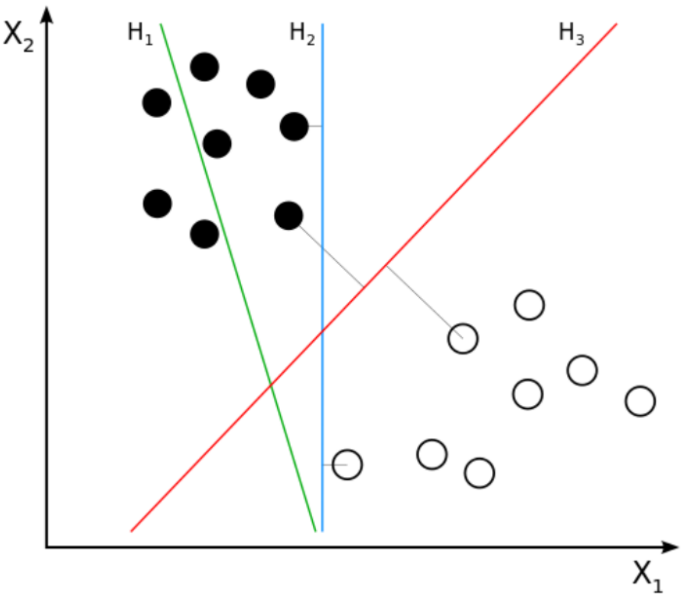


Fig. 2. SVM used to classify points in two dimension

a result, when noise is introduced, these points that are close to H2 may be classified to the other side. In contrast, H3 is much better at dividing points when noise is introduced, so we think H3 is better as a classifier. SVM can be divided into two categories: linear SVM and nonlinear SVM, and their differences are as follows:

- Linear SVM: Linear SVM can be used when all points can be directly divided into two categories by the hyperplane, such as in a two-dimensional plane , if two types of points can be divided into two groups by a straight line.
- Linear SVM: In many cases, data points cannot be directly divided into two categories by a hyperplane, in these cases we need to use some advanced techniques such as kernel to divide data points into two categories.

If the data points are n-dimensional, the SVM uses an (n-1)-dimensional hyperplane to separate these data points into two classes [6,]. For a specific set of data points, there may be multiple hyperplanes that can complete the classification task, but there is a maximum margin hyperplane, which can make the distance from the point closest to the hyperplane to the hyperplane to be the largest(has the maximum distance from both the classes).

So how to find the maximum margin hyperplane is very important in SVM. First we must be able to decide if a point X is positive or negative. The rules for classifying point X using hyperplanes are as follows [6,]:

1) find the normal vector $\hat{W}$ of the hyperplane passing through the origin point O
2) then find the projection of the vector $\hat{OX}$ to the normal vector $\hat{W}$
3) if the length of the projection is less than the length of the normal vector, the point X is identified as negative, otherwise it is identified as positive.

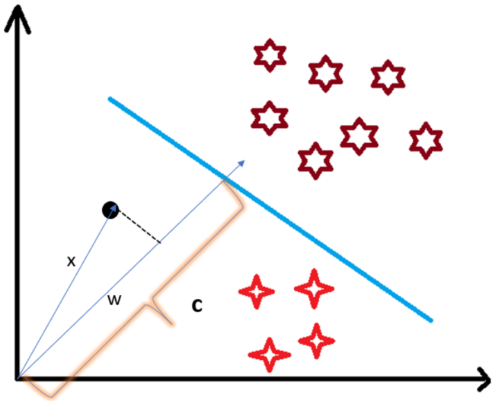


Fig. 3. How to judge a point is positive or negative

As we can define a hyperplane by means of normal vectors $\hat{W}$ and intercepts $\hat{b}$. Suppose we can find parallel hyperplanes on the condition that all points can be divided with these two parallel hyperplanes, and the margin between these two hyperplanes is the largest (The distance between the two hyperplanes can be generalized by the distance of two parallel straight lines: $margin = \rho = \frac{2}{||W||}$).

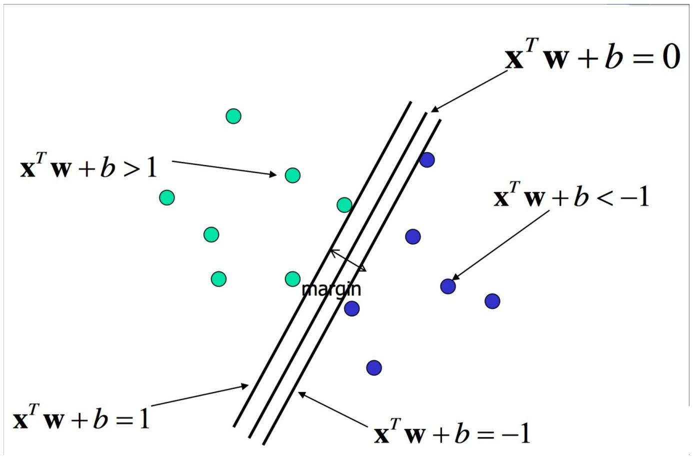

$\mathbf{x}^T\mathbf{w}+b=0$
$\mathbf{x}^T\mathbf{w}+b>1$
$\mathbf{x}^T\mathbf{w}+b<-1$
margin
$\mathbf{x}^T\mathbf{w}+b=1$
$\mathbf{x}^T\mathbf{w}+b=-1$

Fig. 4. maximize margin hyperplane

- For hard-margin SVM, the mathematical expression of the interval maximization problem is that [7,]

$$\begin{aligned} &\min_{W,b} J(W) = \min_{W,b} \frac{1}{2}\|W\|^2 \\ &\text{s.t.} \quad y_i\left(X_i^T W + b\right) \geq 1, i = 1, 2, \ldots n. \end{aligned}$$

  the optimal hyperplane $\hat{W}, \hat{b}$ can be obtained by solving the above equation.
- For soft-margin SVM, the mathematical expression is

$$\min_{W,b} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n} \max\left(0, 1 - y_i\left(X_i^T W + b\right)\right), C \geq 0$$

C is called the penalty parameter. The smaller the value, the smaller the penalty for misclassification. The larger the penalty for misclassification, the greater the penalty for misclassification. When it takes positive infinity, it becomes a hard interval optimization.

For nonlinear datasets, although it is not possible to find hyperplanes which directly to classify them. With the help of *kernel trick* we can convert low-dimension points to high-dimension points, and then find a hyperplane to classify these high-dimension points. Some of the common kernels for SVMs are as follows:

1) RBF Kernel: the RBF kernel is the most commonly used kernel in SVM for nonlinear datasets, its function is

$$f(X_1, X_2) = e^{\frac{-||(X_1-X_2)^2||}{2\delta^2}}$$

$||(X_1 - X_2)^2||$ is the Euclidean Distance between X_1, X_2.
2) Linear Kernal: the linear kernel is the fastest in all kernels, its math expression is $f(X, X_j) = \sum X_j$ [8,].
3) Polynomial Kernel: the function of polynomial kernel is: $f(X_1, X_2) = (X_1^T X_2 + 1)^d$, d is the degree of the polynomial. If X_1, X_2 are points in two dimension, then we can use polynomial kernel to convert them into 5 dimension points [9,].

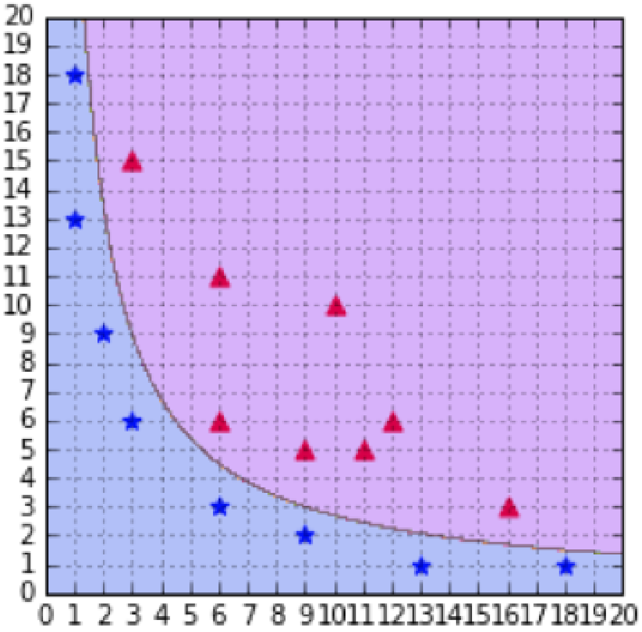


Fig. 5. The SVM using polynomial kernel is able to separate these data

ANN

Water pollution is one of the important environmental problems facing human beings, and its harm is largely caused by the lack of forecasting and early warning and emergency response capabilities. Therefore, it is urgent to build an effective monitoring and early warning system to realize intelligent decision-making and management of water quality. One of the key scientific and technological problems to be solved. Water quality prediction is the pre-link of water quality early warning. [10,]
Intelligent algorithms such as Artificial Neural Network (ANN) have been favored more and more, and also provide new opportunities for solving the problem of water quality early warning (Luger, 2005). Among them, ANN is known for its powerful learning ability and general chemistry has become the focus of research in academia and industry There are many classifications of ANN, among which the three main types used in the field of water quality early warning include feedforward neural network (FFNN), recurrent neural network (RNN) and convolutional neural network (CNN) [11,]. Their characteristics are shown below:

- A neural network is called FFNN if its neuron connections exist only between the input, hidden and output layers. Among them, BPNN uses the BP algorithm for training, which is the most common FFNN. [12,]
- Compared with the FFNN neural network, the RNN neural network contains a circular information flow in its processing unit. Due to the existence of the special structure of circular information flow, RNN has the ability to include the state of the previous moment. The disadvantage of RNN neural network is that when the sequence is too long, RNN is prone to the problem of gradient disappearance or gradient explosion. [13,]
- CNN transmits information through convolutional layers. The sequence composed of multiple convolutional layers will gradually move from the input layer to each output layer for feature extraction, and finally the extracted features are weighted and summed to calculate the result. Compared with FFNN, The neurons in each feature extraction layer of CNN are only connected to the corresponding part in the previous input layer, which significantly reduces the training time and reduces the possibility of overfitting [14,]

C. Deep Learning

Deep learning is a branch of machine learning, and its development is based on the study of neural networks. The feature extraction of traditional machine learning mainly relies on manual work. Manual feature extraction is simple and effective for specific simple tasks, but it is not universal. The feature extraction of deep learning does not rely on humans, but is automatically extracted by machines. There are 4 kinds of common deep learning algorithm: CNN, RNN, GANs and RL.

1) LSTM: LSTM is short for Long short-term memory, it's a special kind of RNN. It's mainly used to solve the problem of gradient disappearance and gradient explosion during long sequence training. To be brief, LSTM can perform better in longer sequences than ordinary RNNs. It has a wide range

of applications in NLP, speech recognition, and time series related fields.
The traditional RNN has the problem of gradient vanishing due to the long sequence, which will cause the RNN to fail to learn long-term dependencies. The gradient vanishing problem means that in the process of RNN model training, some of its parameters will stop changing because the learning process is too long, which makes the training of the model stagnant. So traditional RNN cannot learn long-term dependencies.

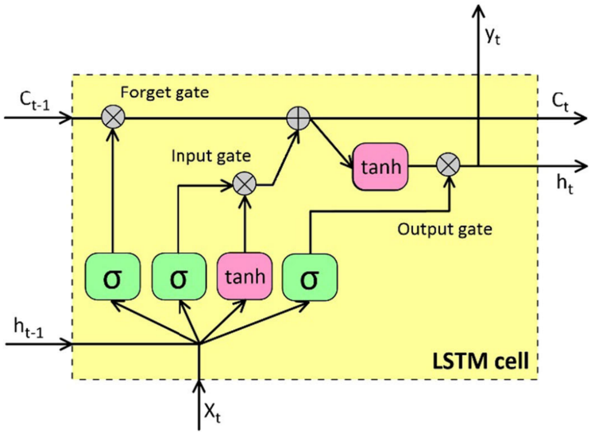


Fig. 6. structure of SLTM neural

In Fig. 6, the green parts on the left below part are the input of neural network, it controls the three multiplicative units: the input, forget and output gates [15,]. As shown in the pink part, these parts are always isolated from the outside world, obviously as LSTM memory or the existence of the main plot. The right uppper corner is the output of neural network [16,]. The main math equation in LSTM is shown below:

1) forget gate logic:

$$f_t = \sigma\left(U_f h_{t-1} + W_f x_t\right)$$
$$k_t = c_{t-1} \odot f_t$$

2) input gate logic:

$$i_t = \sigma\left(U_i h_{t-1} + W_i x_t\right)$$
$$g_t = \tanh\left(U_g h_{t-1} + W_g x_t\right)$$
$$j_t = g_t \odot i_t$$
$$c_t = j_t + k_t$$

3) output gate logic

$$o_t = \sigma\left(U_o h_{t-1} + W_o x_t\right)$$
$$h_t = \tanh\left(c_t\right) \odot o_t$$

LSTM essentially introduces a kind of gating logic inside the neuron, which controls the transmission state depends on the state of gates. It can remembers the information that needs to be remembered for a long time, and forgets the unimportant information. Fig. 7 shows the structual difference between traditional RNN and LSTM. In anomaly detection applications, LSTM has been provd to have very good performance. Due to the characteristics of the SLTM neural node, it is supposed to be able to detect contextual anomalies, and the goal of anomaly detection can be achieved by comparing the expected results with the output of the actual model. [17,]

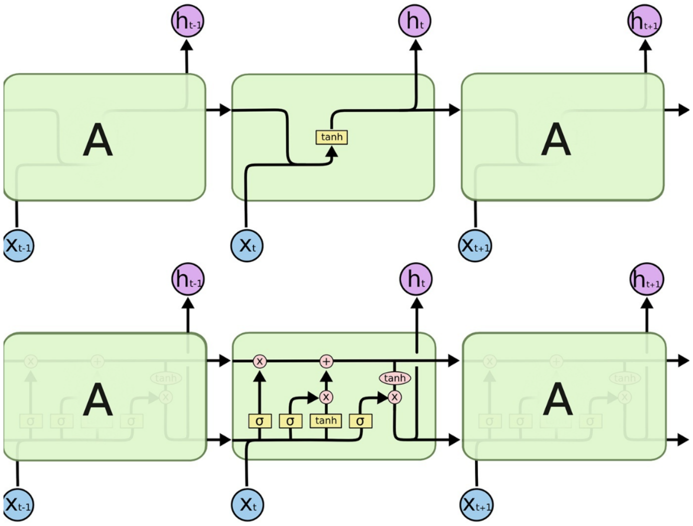


Fig. 7. difference between traditional RNN and LSTM

D. Water data detection method chosen criteria

One of the task of water resource management is water quality monitoring, in order to predict the water quality we can watch two main variable in water. One of them is dissolved oxygen and the other is chlorophyll. Using SVM and SLTM we can build models to predict these variables to monitor the water quality. [15,]
Amony all of the three method: SVM, ANN and LSTM, LSTM has the most accurate prediction for long series of sequential data, because LSTM has the ability to remember or forget the data in an efficient manner than SVM and ANN. So when we are trying to do anomaly detection of water dataset which uses data for many years, it's advised to use LSTM for better prediction [18,].
As for SVM and ANN, SVM is derived entirely on the basis on some several simple mathematical techniques like the partial derivative, and Lagrange multipliers. The inner process is a optimization to find the maximum margin while trying to classify points into two parts. The ANN forms its hyperplane by minimizing the loss function using backpropagation of errors. The hyperplane of ANN can do the job of classify points into two parts but it's not what ANN directly trying to do [19,]. So SVM performs better when trying to find the abnormal dataset from the whole water dataset. Besides, SVM performs better on structural data, and the water data collected during years are structural data.

III. DATA & RESOURCES

The goal of this paper is to use a suitable machine learning algorithm to classify data for detection and identification of water quality. The source of the data is the water quality information obtained by real-time monitoring, which is time series data. Table 1 is an overview of the data. Figure 1 is the graph drawn from the complete data.

In order to monitor the water quality, the Thüringer Fernwasserver-sorgung performs measurements at significant points throughout the whole water distribution system, in particular at the outflow of the waterworks and the in- and outflow of the water towers[23]. Each sample contains six features: the temperature of the water, the pH value, the electric conductivity of the water, the turbidity, the spectral absorption coefficient of the water and the PFM value of the sensor panel. Besides, each sample has on label data: Event (”True” or ”False”), which is artificial marked to show the water quality has changed significantly or not. Our goal is to predict the Event based on the 6 features described above.

There is a total of 132,480 samples, and it should be noted that among the 132,212 non-missing data, only 209 data (0.1581%) correspond to the Event being ”True”, which means that the data is highly imbalanced. If we do not use any machine learning algorithm for training, and directly predict all the sample labels as ”False”, then the prediction accuracy will be as high as 99.8419%. In this case, the accuracy rate will lose its evaluation value, and we need to find a more suitable model evaluation metric.

TABLE I
DESCRIPTION OF THE DATA

Column name	Description
Time	Time of measurement, given in following format: yyyy-mm-dd HH:MM:SS
Tp	The temperature of the water, given in °C.
pH	pH value of the water
Cond	Electric conductivity of the water, given in S/m
Turb	Turbidity of the water, given in FNU
SAC	Spectral absorption coefficient, given in 1/m
PFM	Pulse-Frequency-Modulation, given in Hz
EVENT	Marker if this entry should be considered as a remarkable change resp. event, given in boolean.

IV. METHODS & EXPERIMENT DESIGN

A. Data Preparation

We will do some simple processing to the data. Measuring the six features requires the use of six different devices, but various real-world complexities can lead to device failure or missing data storage. If one of the machines fails, it will affect the reliability of the entire sample, and we will deal with these missing data. Figure 2 shows the missing proportions corresponding to the six features. It can be seen that for each feature, the missing ratio is less than 0.2%. If we define: For a sample at a certain time point, if at least one feature is missing, then the entire data is a missing sample. According to this definition, it can be calculated that the proportion of all missing samples (1242) to the total number of samples (132480) is 0.9375%.

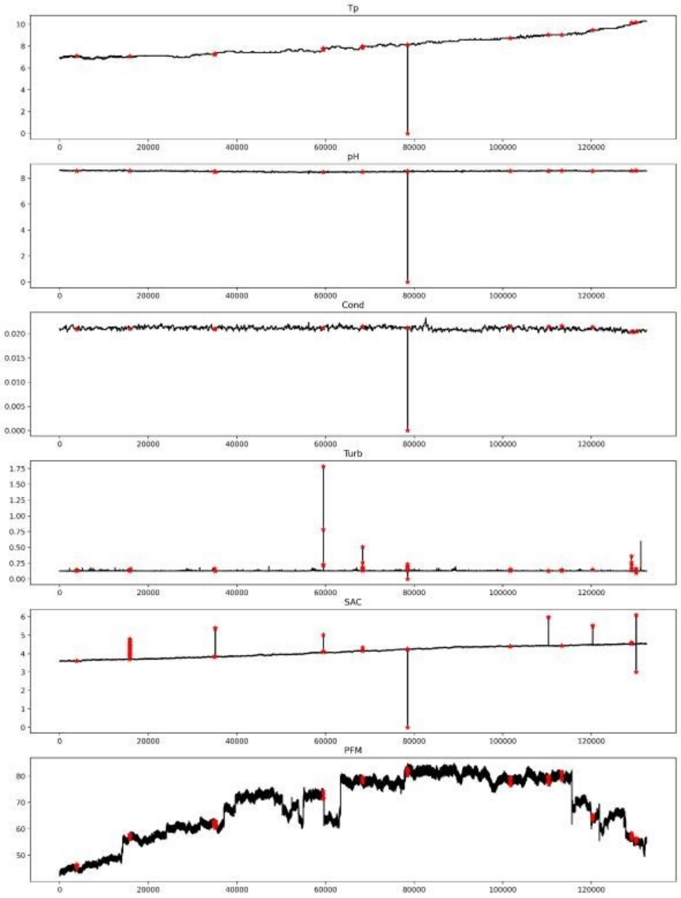


Fig. 8. The graph drawn from the complete data, the Event corresponding to the black line is ”False”, and the Event corresponding to the red data point is ”True”.

The missing processing of time series data should ensure its time continuity. Commonly used missing data processing methods are: using the mean value, the following value, the previous value to fill in the missing values, or directly delete the entire sample. As can be seen from Figure 1, for the data features of this article, the values of some features corresponding to the samples whose Event is ”True” will be very large. Therefore, it may not be a good choice to use the mean to fill in the missing values. We can fill in the missing values with the following value, the previous value, or simply delete the entire sample. Since the proportion of missing samples in this paper is extremely small, only 0.9375%, the missing samples can be deleted directly, and its impact on the time series can be ignored, and the introduction of new uncertainties by other methods can be avoided.

B. Feature selection

The selection of features has a great impact on the performance of machine learning, so in this section we will focus on the selection of features. There are two main purposes: 1. To observe whether there is a very simple feature selection that makes the two types of Events (”True” and ”False”) easily distinguishable. This means artificially extracting key features, making the data easier to classify. 2. Calculate the correlation between features. If some features are highly correlated, it means that some of them may be redundant, which will

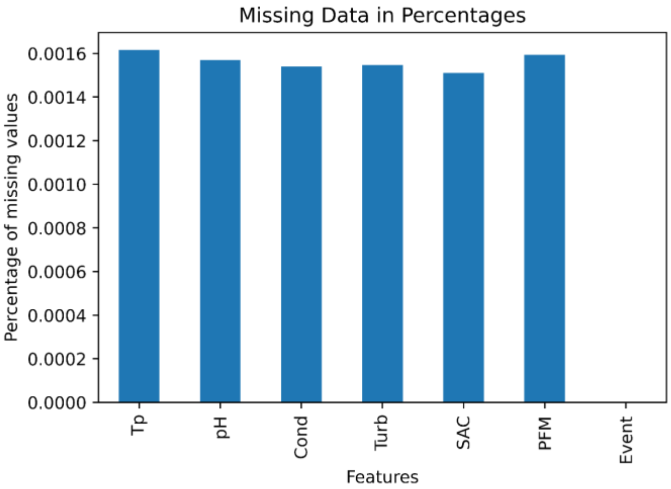


Fig. 9. Missing Data of Each Feature in Percentage

affect the training effect of machine learning and reduce the accuracy.

In order to achieve the first purpose, we have six features, select two features at a time, draw them as a scatter plot, and set the two events (”True” and ”False”) to two colors, see Figure 3. First, it is very obvious from the figure that there does not seem to be some kind of simple feature selection so that the two classes can be very easily distinguished. Second, a very interesting phenomenon can be observed: most of the data far from the center of the cluster is red, that is, which means the Event are ”True”. This corresponds to the data change characteristics in Figure 1. Some characteristic values of data points with changes in water quality will change significantly. Finally, it should be pointed out that this method of data presentation is very limited: it is limited up to two dimensions, so it cannot show higher-dimensional relationships. Besides, it cannot reveal nonlinear relationships between features.

To achieve the second purpose, the correlation coefficients between the various features are calculated. The correlation coefficient actually reflects the degree of linear correlation between features. The results are shown in Figure 4. The color of the graph reflects the magnitude of the correlation coefficient intuitively. What we need to pay attention to are the correlation coefficients of off-diagonal positions, they represent ”correlation between features”. It can be seen that only TP-SAC has a high correlation, and the rest of the correlation coefficients are all less than 0.5. This means that all features are preserved and there is no redundant data.

C. Evaluation Metrics

In the data description section, we have introduced that there is a natural quantitative imbalance in the water quality measurement data, that is, only 0.1581% of the sample events are ”True”, and the rest are ”False”. This means that overall prediction accuracy has lost its evaluation value and we need metrics that can accurately evaluate models on imbalanced datasets.

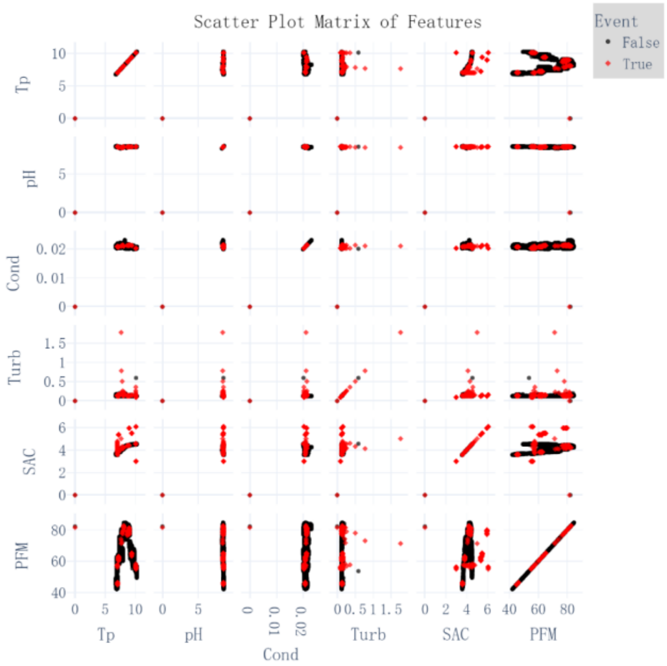


Fig. 10. Scatter Plot Matrix of Features

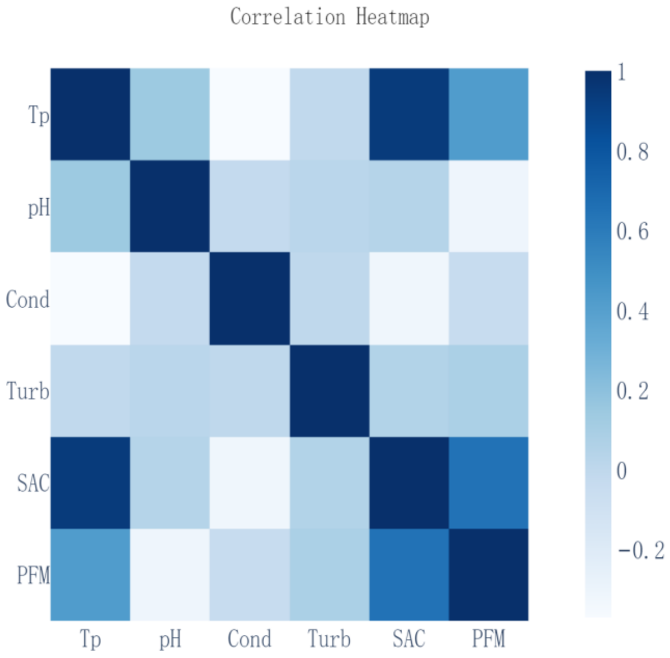


Fig. 11. Correlation Heatmap

First, we will introduce the concept of confusion matrix. As shown in Figure 5, this is a schematic diagram of the confusion matrix of binary classification, where TP, FN, FP and TN are the number of samples of the corresponding category. The confusion matrix can represent the prediction results very intuitively. Next, the true positive rates and false positive rates can be calculated:

$$TPR = \frac{TP}{TP + FN}$$

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

The prediction result of a classification problem is usually a probability score between 0 and 1, and we can set a threshold, and when the score is greater than this threshold, the prediction result is positive (or negative). It means that we can get different confusion matrices by adjusting the threshold, and then calculate TPR and FPR. Taking FPR as the abscissa and TPR as the ordinate, and connecting the points corresponding to different thresholds, the ROC curve can be obtained. The area under the ROC curve is called AUC (Area Under ROC Curve), which can be used for model evaluation.

Now we introduce the calculation formula of F1 score. First calculate the Precision, which means that in the predicted positive samples, the proportion of accurate predictions:

$$Precision = \frac{TP}{TP + FP}$$

Then calculate Recall, which represents in all real positive samples, the proportion of samples that are accurately predicted:

$$Recall = \frac{TP}{TP + TN}$$

Finally, the F1 score can be calculated based on Precision and Recall:

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 score is the harmonic mean of the Precision and Recall, [24,] with a numerical value ranging from 0 to 1, it can be a good estimate of the prediction accuracy of imbalanced samples, the closer the model is to 1, the better.

V. RESULT

Divide the dataset into training and testing parts, then use the model described in the previous section to train on the training set, make predictions on the test set, and compare with the true labels. The results are presented below.

The first result is Figure 6, which is the confusion matrix of the 5 methods. He has a very remarkable feature, the numerical value (TN) in the upper left corner is very large, which is caused by the uneven distribution of the data set. The ROC curve is shown in Figure 7. It can be seen that the coverage area of the five algorithms is relatively large, and the SVM (Linear) performs the worst, which means that the linear separability of the data is poor. Sort by AUC from large to small: $LSTM > SVM(RBF) > ANN = SVM(POLY) > SVM(Linear)$.

Next, the evaluation indicators are calculated according to the confusion matrix, and Table 2 shows the calculation results. Sort by F1 score from large to small: $LSTM > SVM(POLY) > SVM(RBF) > ANN > SVM(Linear)$.

Overall, LSTM has the best overall performance. The possible reasons are: water quality data is time series data, and LSTM is a time series algorithm, which can take into account long-term dependencies. This may mean that some insignificant changes are captured by the LSTM before the Event of water quality turn into ”True”. However, ANN and SVM do not consider the data as time series data during training, so they cannot capture these changes, which leads to worse performance.

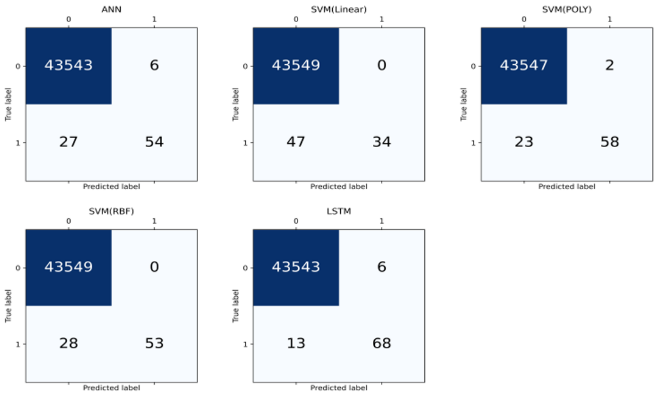


Fig. 12. Confusion matrices of 5 methods

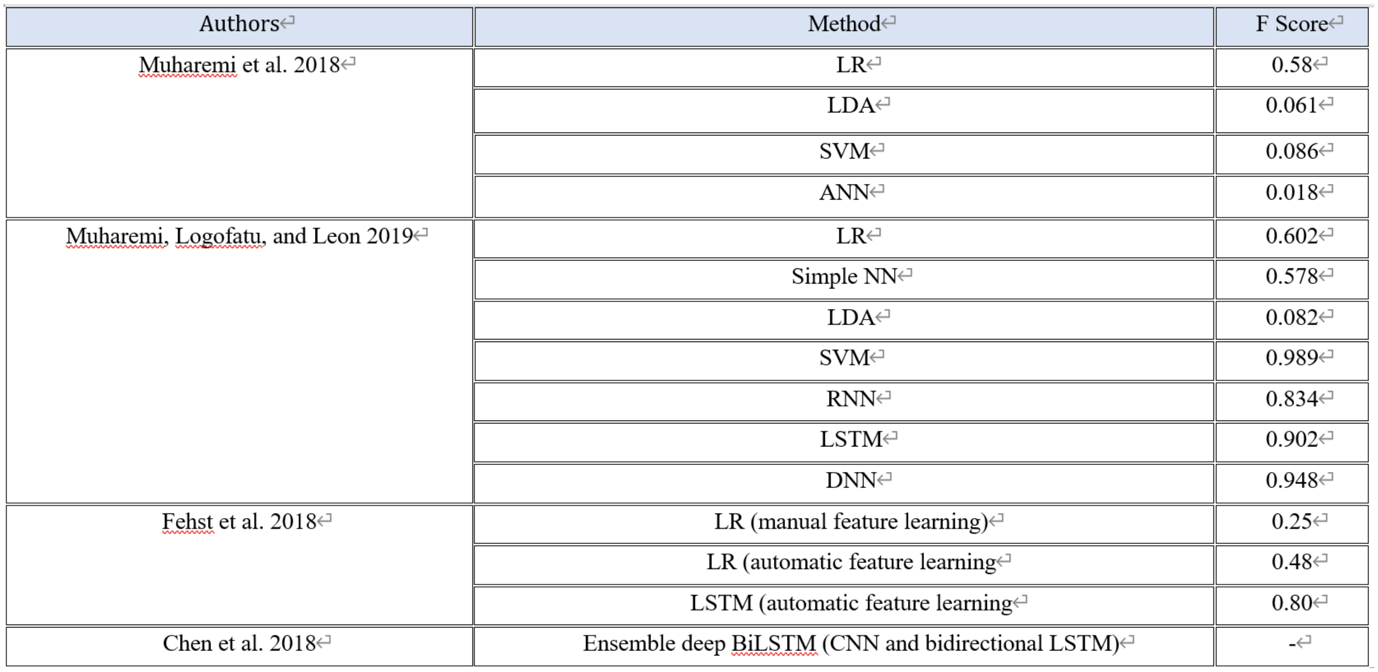
VI. RELATED WORK AND DISCUSSION

The goal of this paper is to use suitable machine learning algorithms for the classification of water quality time series data.

After data preparation, feature selection, and machine learning training, we used ROC, AUC, and F1 score for a comprehensive evaluation, and finally found that LSTM performed better than SVM and ANN. The results of this paper are then compared with those published by other researchers, and several topics are further discussed, including weaknesses and limitations of this work, directions for improvement of our research, and some considerations for real-world applications.

First, compare our findings with those published by other researchers. Eustace M. Dogo et al. systematically investigate the application of machine learning for monitoring abnormal water quality data[25]. They point out that, in general, deep learning performs better than ML. This is consistent with our findings. In addition, in table3, Valerie Fehst et al. used three ways to classify the water quality data of the GECCO 2018 Industrial Challenge. The first way is to directly use logistic regression to classify the data; the final F1 score is 0.25. The second method is to artificially select features: lag operator c, integrating the features, complexity of the probability distribution, and composition gradient. After artificially selecting features, combine the original features

TABLE II
SUMMARY OF RELATED WORK



Authors	Method	F Score
Muharemi et al. 2018	LR	0.58
	LDA	0.061
	SVM	0.086
	ANN	0.018
Muharemi, Logofatu, and Leon 2019	LR	0.602
	Simple NN	0.578
	LDA	0.082
	SVM	0.989
	RNN	0.834
	LSTM	0.902
	DNN	0.948
Fehst et al. 2018	LR (manual feature learning)	0.25
	LR (automatic feature learning	0.48
	LSTM (automatic feature learning	0.80
Chen et al. 2018	Ensemble deep BiLSTM (CNN and bidirectional LSTM)	-

to form a new dataset, and finally use logistic regression for classification. The F1 score was 0.48. The last method is consistent with this paper, using LSTM with automatic feature extraction to directly train the original data, and the final F1 score is 0.80. LSTM performs the best, which is consistent with our research. They pointed out in the article that the data of automatically extracted features can reflect nonlinear information, so the results of LSTM and logistic regression with artificially selected features perform better than the logistic regression with raw data. This may partly explain why our study's SVM (Linear) performed the worst since this method cannot reflect nonlinear information.

It is worth mentioning that Muharemi et al. also conducted related research on water quality anomaly detection in table 3, and the data set they used was the water quality data of the GECCO 2017 Industrial Challenge. Unlike our data, the GECCO 2017 data provides 9 features and one label, the 9 features are Tp, Cl, pH, Redox, Leit Trueb, Cl_2, Fm, and Fm_2. The machine learning models they use are SVM, DNN, LSTM, RNN, LogRegression, Simple NN, and LDA. The final results show that the SVM model performs the best, with an F1 score of 0.9891, while the F1 score of LSTM is 0.9023. And a very interesting phenomenon is that they used SVM to win this competition, and after the competition, when they applied the trained model to brand new data, the performance of the model became very bad, and the F1 score of the SVM model Changed from 0.9819 before to 0.36. They concluded: 'All models are wrong; some models are useful. This shows that real-world complex problems are difficult to solve all at once. The performance of LSTM is better than that of SVM, which is contrary to our findings, and the possible reasons are:

1) The datasets are different. The GECCO 2017 data provides 7 features and one label, while our study uses the GECCO 2019 data with only 6 features and one label. It is reasonable that different machine learning models perform differently on different datasets.
2) They used a trick that helped them win the competition: in addition to the original data, they also used artificially selected new feature terms like the study[26], i.e. some interaction terms. Unfortunately, they did not elaborate on the specific method in their article. However, our study only used the original 6 features, which may cause differences in the performance of machine learning algorithms.
3) The F1 score of the SVM model they showed is the best among many SVM models with different kernels and parameters, but our study only considers the SVM models of three kernels: Linear, POLY, and RBF. There are other kernels like Gauss that we didn't consider. Moreover, our research did not use the technique of tuning parameters but directly trained the results using the built-in training parameters of Sklearn. This may also be the reason why our SVM models are worse than theirs.

There is also a topic worth discussing: the treatment of unbalanced samples. In our study, the proportion of samples whose Event is "True" is only 0.1581%, in order to accurately evaluate the performance of machine learning models on unbalanced samples, we combined ROC, AUC, and F1 score. This evaluation is only done during the testing phase, we do not take any measures during the training phase. Unbalanced samples will affect the decision boundary of the classification algorithm, because, for a small number of samples of a certain type, the machine learning algorithm may not fully learn their characteristics, which will affect the final performance. One solution is to use the resampling method to generate many new samples to balance the class proportions. Eustace M. Dogo used eight resampling methods: RUS, Tomek, RENN, ROS, SMOTE,ADASYN, SMOTE+Tomek, and SMOTE+ENN[27]. Finally, it is found that using SMOTE combined with DNN can be very effective in improving model performance. This is something that our study did not take into account and could be an improvement direction for future research work.

In addition, for an excellent machine learning detection system, after it is actually used, it should constantly update its parameters according to new data, so as to explore the value of the new data and improve the recognition accuracy.It is pointed out that although Muharemi's model won the championship in the GECCO 2017 competition, after the competition when the trained model is applied to completely new data (meaning not present in the training data and hyperparameters), the performance of the model becomes very poor, the F1 score of the previous best model SVM changed from 0.9819 to 0.36[28]. This shows that the results learned from the smaller datasets are hardly sustainable in the real world, and we need to continuously train the model with new data to improve the adaptation of the model to the real world. The data span used in this article is only three months, with a total of 132,480 samples, so the training cost is very low. The training time of the three machine learning methods does not exceed 5 minutes, but LSTM is longer than SVM and ANN. (The main reason is that LSTM has more parameters to control input, forget and output, which means that it will take more training time to converge.) But in real application scenarios, there will be more and more data as time accumulates. If a system has been used for two years, it can be estimated that the total number of samples will reach more than 100,000. This creates challenges for updating machine learning models, as the larger the amount of data, the longer the training time. From this point of view, although LSTM has the best classification performance, its training time is longer than that of SVM and ANN. If real-world applications require rapid model updating, LSTM is not necessarily the best choice. Perhaps we can use data dimensionality reduction algorithms such as Principal Component Analysis (PCA) to reduce the dimensionality of the data to increase the training speed, but doing so may reduce the F1 score, which needs further research.

Finally, we have to ask a question: Whether it is the work of this paper or the work of other researchers, it is an attempt to achieve: when the system detects a new set of data samples, it is fed back to the machine learning model, and the model will tell you if the water quality has changed abnormally. However, such anomaly changes are ”already happened” because current machine learning models can only predict changes based on data on ”changed” water quality. However, changes in water quality in the real world may cause a lot of losses, such as human poisoning after drinking, ecosystem damage, and animals being unable to survive, etc. We do not want to see these losses. Is there a detection system that can tell us before the water changes: abnormal changes in water are about to occur, please check for water-related human or natural activities? In this way, the detection system can remind us to do preventive work before the water quality changes, so as to prevent the water quality from changing as much as possible and reduce the possible losses. A possible method is: there are only two types of data now, one is normal (Event is ”False”), and the other is abnormal (Event is ”True”), maybe we can add a new category for a period of time before ”True” (how long this period of time needs to be carefully considered), and the data of this period is manually marked, indicating that the water quality ”will” change abnormally. This presents a more difficult challenge to our research work, but it has greater real-world application value.

VII. Conclusion

The purpose of this paper is to use machine learning algorithms to identify changes in water quality. The source of the data is the GECCO 2019 Industrial Challenge, which is time-series data with six detected water quality features and one label. First, we carried out data preprocessing to remove a very small proportion of missing samples. Then, the selection of features was carried out, and it was found that the correlation between features was weak and the data could be directly used for machine learning training. We then use three machine learning algorithms for training and prediction: ANN, SVM, and LSTM. Since the distribution of data samples is extremely unbalanced: the proportion of positive samples is only 0.1581%, we comprehensively use confusion matrix, ROC curve, AUC and F1 score to evaluate the model. In the end we found that LSTM performed best on the classification of this time series data, and the F1 score is 0.88. The work of this paper can be applied to real-time monitoring of water quality, which is beneficial to protect the living environment and health of human beings.

VIII. Declarations

Declaration of Originality

I am aware of and understand the University of Exeter's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices.

Declaration of Ethical Concerns

This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also no security or safety critical activities have been carried out.

References

[1] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, “A unifying review of deep and shallow anomaly detection,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021. II-A, II-A, II-A

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882 II-A

[3] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *Ieee Access*, vol. 9, pp. 78 658–78 700, 2021. II-A

[4] P. Gogoi, D. Bhattacharyya, B. Borah, and J. K. Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification," *The Computer Journal*, vol. 54, no. 4, pp. 570–588, 03 2011. [Online]. Available: https://doi.org/10.1093/comjnl/bxr026 II-A

[5] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nature Reviews Molecular Cell Biology*, vol. 23, no. 1, pp. 40–55, 2022. II-B

[6] K. Vos, Z. Peng, C. Jenkins, M. R. Shahriar, P. Borghesani, and W. Wang, "Vibration-based anomaly detection using lstm/svm approaches," *Mechanical Systems and Signal Processing*, vol. 169, p. 108752, 2022. II-B, II-B

[7] Q. Ma, C. Sun, B. Cui, and X. Jin, "A novel model for anomaly detection in network traffic based on kernel support vector machine," *Computers & Security*, vol. 104, p. 102215, 2021. II-B

[8] B. Wang, X. Zhang, S. Xing, C. Sun, and X. Chen, "Sparse representation theory for support vector machine kernel function selection and its application in high-speed bearing fault diagnosis," *ISA transactions*, vol. 118, pp. 207–218, 2021. 2

[9] H. Nguyen, X.-N. Bui, Y. Choi, C. W. Lee, and D. J. Armaghani, "A novel combination of whale optimization algorithm and support vector machine with different kernel functions for prediction of blasting-induced fly-rock in quarry mines," *Natural Resources Research*, vol. 30, no. 1, pp. 191–207, 2021. 3

[10] M. Wu, W. Zhang, X. Wang, and D. Luo, "Application of modis satellite data in monitoring water quality parameters of chaohu lake in china," *Environmental monitoring and assessment*, vol. 148, no. 1, pp. 255–264, 2009. II-B

[11] G. F. Luger, *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education, 2005. II-B

[12] V. Nourani, M. T. Alami, and F. D. Vousoughi, "Self-organizing map clustering technique for ann-based spatiotemporal modeling of groundwater quality parameters," *Journal of Hydroinformatics*, vol. 18, no. 2, pp. 288–309, 2016. II-B

[13] F. Gers, "A., jrgen. schmidhuber, and fred. cummins. 2000," *Neural Computation*, vol. 12, pp. 2451–2471. II-B

[14] P. Dhruv and S. Naskar, "Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): a review," *Machine learning and information processing*, pp. 367–381, 2020. II-B

[15] R. Barzegar, M. T. Aalami, and J. Adamowski, "Short-term water quality variable prediction using a hybrid cnn–lstm deep learning model," *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 2, pp. 415–433, 2020. II-C1, II-D

[16] J. Zhou, Y. Wang, F. Xiao, Y. Wang, and L. Sun, "Water quality prediction method based on igra and lstm," *Water*, vol. 10, no. 9, p. 1148, 2018. II-C1

[17] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using lstm networks," *Computers in Industry*, vol. 131, p. 103498, 2021. II-C1

[18] S. K. Lakshminarayanan and J. P. McCrae, "A comparative study of svm and lstm deep learning algorithms for stock market prediction." in *AICS*, 2019, pp. 446–457. II-D

[19] J. Ren, "Ann vs. svm: Which one performs better in classification of mccs in mammogram imaging," *Knowledge-Based Systems*, vol. 26, pp. 144–153, 2012. II-D