Water Quality Classification Using Machine Learning

N. Raviteja[1], N. Saiteja[2], N. Sreenu[3], G. Senthilvelan[4], Dr. D. Usha[5], Dr. T. Kumanan[6]

Final Year B.Tech of CSE[1,2,3], Asst Professor[4],Professor[5,6]
[1,2,3,4,5,6]Department of CSE, Dr.MGR Educational and Research Institute
natakarani.raviteja1432@gmail.com,nsaiteja148@gmail.com,nakkasreenu6@gmail.com

ABSTRACT

The quality of water has been significantly influenced by many other pollutants over the past few years. It has the direct impact on human health and environment. The WQI works as an indicator of water management, efficiency. Knowing the quality of water and even how to model the quality in prediction benefits the war against water pollution. The objective of the study is to establish a reliable prediction model for river water quality, that is able to identify the index value is based on the river water quality standards. In this project, the project Inspect and compare the performance of many classification models and algorithms to find which attributes were prominent in classifying river water quality. A total of eleven sampling stations, spread across different points on the River flowing through Kerala and Tamil Nadu, have been chosen for the data collection. The water quality index is measured by 7 different environmental factors affecting the quality of water, including the dissolved oxygen level, temperature, pH, hardness, chloride etc. Supervised Machine learning algorithms such as logistic regression, support vector regressor have been used to develop a model for predicting water quality. A classification model was developed using SVM classifiers, SVM, to classify water quality index. Logistic regressor efficiently predicts water Quality index, SVM classifier classifies water Quality index with an accuracy of 83%. The built models presented favourable results regarding water quality index predictions and classification.

Keywords: River Water Quality, Water Quality Index (WQI), Machine learning, SVM**,** Logistic Regression, SVC, SVR.

I. INTRODUCTION

To human beings needs water as the most critical fundamental requirement for life since living organisms need water to sustain their life on earth. Although it hardly has been an object of science fiction, better quality and quantity in the amount of available water on our planet is more important to life on earth than any of the other things we may be able to think about. When the pollution level is moderate, it's common for water-dwelling species to inhabit the area. But when the pollution rises, the oxygen levels in the water decline, resulting in dire consequences.About the quality of environmental water sources such as lakes, rivers, and streams, a high proportion of them has standards that prove their value. Guidelines ares applicable not only to bodies of water of all types for all applications and uses It does not have to be so salinity nor.

This harms the plant or the soil therefore distorting the entire ecosystem. Industrial applications also have to have different types of water quality as particular processes are of a different nature
mainly for the consumption of humanity. The cheapest methods of getting freshwater like ground, and surface water are the natural waters resources. The human as well as the industrial activities and other environmental actions contaminate natural resources. For instance, irrigation water should not have excessive salinity or contain poisonous substances which can be transferred to plants or soil, thereby destroying the ecosystems.

Industrial water particularly suitable for commercial uses, different properties are desired depending on the nature of the industrial processes. Some of the sources of fresh water which are considered to be low-priced include ground and surface water, they are classified as natural water resources. Yet that sources can be contaminated by human?industrial activities and other natural processes. The spoiling of water quality at an alarming pace has been caused by the fast industrial development as a result. Even so, infrastructures, unaccompanied by public consciousness, and less clean features considerably alter standards of drinking water. In reality, hazards brought about by polluted water for drinking are so threatening to the health and that of the environment and infrastructures used at large.

According to the World report, some 1.5 million people die annually, due to illnesses caused by diseases related to poor sources of water. It is obviously noted that in developing countriesthe nations 80% of health conditions are as a result of the contaminated water. Each year results in an estimate of five million deaths and 2.5 billion people with diseases. It is advocated not to forget the temporal measurement for anticipating the Water Quality (WQ) types to avoid missing the seasonal differences in the WQ. But the combined model dissimilarity is superior to that of using a single model for prediction of WQ. Methodologies have been proposed that make the prediction and modelling of the WQThe rapid progress of industrialization has resulted in a marked deterioration of water quality. Inadequate infrastructure, lack of recognition, and unsanitary practices greatly contribute to the decline in the quality of drinking water. This contaminated water poses a serious threat to public health, presenting a multitude of long-term consequences that not only affect our well-being but also harm the environment and our infrastructure. Recent studies conducted in the United States provide evidence to support this alarming issue.

Annually about 2 million people do not receive the deserved outcome of their situation the reason of which can be different. The problem of improper water supply systems contributes to that the figures of severe health conditions, in developing countries, make almost 15 million. Every year around 2.5 billion people adds up to the cases and eventually lose their lives by taking waterborne diseases. To this, statistics show that half a million people annually die from these waterborne illnesses. In human cruelty deaths or even accidents as well as the terrorist attacks are investigated too. Unfortunately, it is not possible to put these innovative techniques out now because it has not been proven that these methods can predict and study water quality (WQ) accurately. As a result, it is essential to explore the temporal characteristics of water quality patterns in view of annual changes not forgetting temporal shifts, so as to understand this critical parameter.

Only one model will not be able to demonstrate the accurate image of water quality; rather two combined models are needed for that purpose. Developing estimate method of water quality along with performance simulation method can be the water quality strategy. Techniques such as statistics, visual modeling, algorithm analysis and predictive algorithms are some of the skills often sought after. Going monthly, as the statistical approaches are all of the multivariate and helpful in determining the correlation and relations of different water quality parameters. Many techniques are included and adduced for example, transitional probability, regressive analysis, geostatistical analysis, multivariate analysis and interpolation. This includes consequences like population rise, extracting organics/chemicals from plants, and industrialization, which have negative impacts on ecosystems around the world. Water quality models are indispensable in framing/elucidating water pollution problems.

Fitting for predicting and modeling water farm, is other of consequential and conclusive models. From a mechanism standpoint, water quality is a feature of this advanced system which is more precise and includes data that any ordinary system would not be able to handle. An all puppose model of this kind would certainly get the attention of any body of water. The purpose of this project is to create a predicting model with high accuracy on the quality of the river water having a framework as a basis for this. The leaching quality of river water can be observed at eleven sepatraterivers locations fed by two states namely Kerala and Tamil Nadu. This will make it possible for the researchers to use statistical analysis of the data in order to reveal the distribution patterns and the cases of correlation.Estimation of water quality index values and classification of water quality indexes are implemented with machine learning algorithms. The metrics applied to evaluate the classification models of the water quality index are accuracy, precision among other details.

II. LITERATURESURVEY

The water pH is using the model and predict modelled on the paper which is called to as predict pH in quality of water. With that, the process which involves WQI and WQC algorithm for calculation; has been introduced here in our high-end technology. Deep Learning algorithms [1] such as NARNENT and LSTM use them. This also includes learning algorithms since these are the basis for ML including SVM, KNN and Naïve Bayes and making the WQI classification. According to their superior robustness they have 6 models from the data set that predicted waterfare. Though this also is approximative with the difference in the regression coefficient when gathered following a compression of testing phase. The models can be estimated further in the future work, so these implement for NP samples also of water quality, where different kinds of water are used.

As a real-time method for anomaly detection from historical patterns of water quality data, Zhang et al. [2]proposed using double time-moving windows to detect anomalous data by the algorithm in live . The auto-regressive linear combination model were used to construct the algorithm discussed above. The algorithms were validated with monthly flow of riparian region PH data tested using real and verified environmental water quality measurement station. Further, from the experimental results, it can be clearly represented that the algorithms every negative case is at a lesser number and has increased anomaly detection performance when compared to AD and ADAM algorithms.

Sillberg et al. [3] have proposed a new method utilizing the attribute-realization (AR) and support vector machine (SVM) system for the classification based on machine learning concerning Chao Phraya river water quality issue. By the linear function, the AR has indicated essential factors for improving cluster quality that means bearing in mind fulfillment of these elements. nitrogen Trihydride, Trichlorobenzene, Biochemical oxygen demand Fluorescein di-beta-D-cellobioside, Dissolved_Oxygen and Sal was the most important subcharacteristics of the categorization in terms of contributed values from 0.80 to0.98 while the other characteristics gave values between 0.25- and zero so these were excluded from further considerations The following configurations best represent themselves as second order PCM The best classification results were obtained using SVM linear approach, with accuracy 0.94 and average for precision 0.84; recall 0.84; F1-score – 0.84 coincided with a relative measure of overlap determined by formula

This section examines the WQI [4] using supervised machine learning algorithms under a single index which summarizes all over quality of water and class signifies the general quality/quality of water. The stated techniques included gradient boosting with a learning rate of 0.1, as well as polynomial regression with a second degree and those turned out to be very effectively in predicting WQI. Later on that specific WQI was evaluated through mean absolute error (MAE) which was found to be 1.9642 and 2.7273 respectively In this scenario, the MLP obtained at (3,7) is outperforming all other networks in terms of accuracy for classification by about 85.07% .

III. MATERIAL AND METHODS

The impact of water quality on human health and the environment is significant, given its diverse applications in drinking, agriculture, and industry. Key parameters like dissolved oxygen (DO), total coliform (TC), biological oxygen demand (BOD), Nitrate, pH, and electric conductivity (EC) govern water quality. The evaluation process involves five stages: data pre-processing with min-max normalization and missing data addressed using Random Forest (RF), feature correlation examination, application of machine learning classification, and assessing the model's feature importance. Following analysis, the model stacking strategy outperformed all individual base models.

Disadvantages of Existing System

- Classification of the data not easy
- Need more dataset values to train model
- Low accuracy

- Comsumption of time and Space is also extremely big level
- A Smallamount of water qualities can be covered
- Procedure is extremely slow

This model emphasizes an proposed method that which is design using the some of the machine learning algorithms. Here are the process to perform using machine learning,SVM the algorithm that separates the dataset by non-linear approach and Here we are using SVC to hyperplane classifies the dataset linearly. This suggests technique useful for tracking vast field of water. There should be solutions for classifying and detecting the water quality to get some knowledge which will later help in improving the quality of water. So, patterns on the water quality will help in identifying what problems it has.

Advantages of Proposed System

- Training with less data set
- Classification of data is Easy
- Easy to detect leakage
- High performance
- Easy Identification

Calculation of Water Quality Index (WQI):

WQI is a standard parameter on which all the researchers are using to evaluate the quality of water for human consumption; for example, El Baba et al., (2020), Reyes-Toscano et al. (2020), Zhang et al., (2020), Maskooni et al. (2020) and Bahir The process starts by assigning weights (1-4) to factors that are regressive in nature. The factors that have received a score of 4 are Mineralization, Cl−, and o − the reason behind that is due to the direct effect variables have on water quality and human health (Ahmed M. I. et al., 2020). Bicarbonates (HCO3−) obtain the minimum value of “ 1 “ .

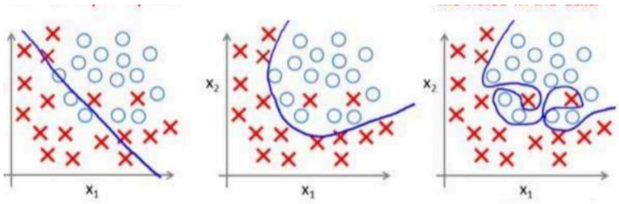


Fig1 Relative weights, assigned weights, and WHO limits are described in table 1

$$\mathrm{Wi} = \frac{\mathrm{wi}}{\sum_{i=1}^{n} \mathrm{wi}}$$

-------(1)

where ‘ Wi ‘ is the relative weight, ‘wi ‘ signifies weight/parameter, and “n ‘ is the parameter count.

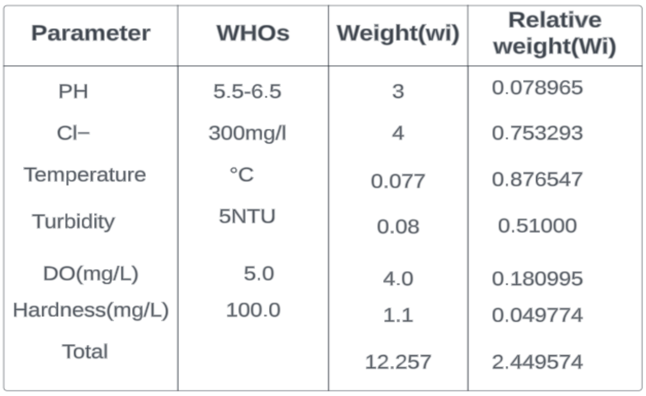
So, by the help of the Eq following qi determinations comes down on every parameter. (2) in which the quality rating scale is “qi”, the chemical concentration of the sample would be “Ci” in milligrams per litre and the WHO’s drinking water quality standard in milligrams per litre would be “Si”.

qi = (Ci/Si) x 100 ---------- (2)

Where “qi” is the quality rating scale .“Ci” is the chemical concentration/water sample (mg/L).“Si” is the WHO drinking water quality standard (mg/L).

In addition , a subindex of the ith parameter is calculated using Eq3
Where

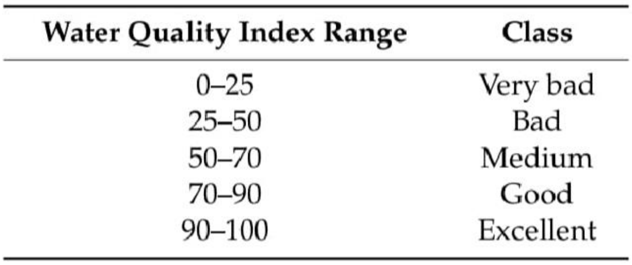
SIi = qi x Wi ----------(3)



Parameter	WHOs	Weight(wi)	Relative weight(Wi)
PH	5.5-6.5	3	0.078965
Cl−	300mg/l	4	0.753293
Temperature	°C	0.077	0.876547
Turbidity	5NTU	0.08	0.51000
DO(mg/L)	5.0	4.0	0.180995
Hardness(mg/L)	100.0	1.1	0.049774
Total		12.257	2.449574

Table 1 Relative weights and Assigned weights of physicochemical parameters

Once we had make an estimate of the WQI, we determined the water quality class (WQC) of each sample usingthe WQI in classification algorithms as shown in Table 2.



Water Quality Index Range	Class
0–25	Very bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

Table 2: Water quality index

SVM

Support Vector Machine (SVM) is one of the most important algorithms which are commonly used as a baseline to compare model performances during the research.” It is a linear model-based family, where training consists of transforming the original vector into a more dimensions space to finding a dividing hyperplane that would have the highest distance. This algorithm builds two orthogonal hyper planes on the either side of the separating hyper plane to maximize the distance between those hyper plates. This means that an estimating a larger gap is accompanied with a lowering the average error of the classifier. The mathematical model of the given SVM algorithm has the input and output data sets X and Y, with x\epsilon X and y\epsilon Y, which aim to learn a X ɔgnisor f(x,\theta) (Consider \theta as the function parameters).

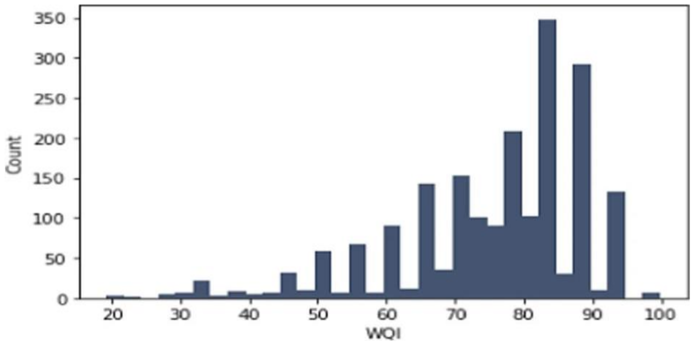


Fig2: Determines the distribution of calculated feature (WQI). The statistical calculation for the feature (WQI)

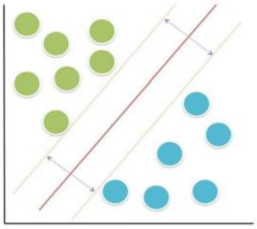


Fig 3 General graphical scheme of SVM algorithm. The model based on two hyperplanes with the maximum distance separates green and blue classes

The representation of the basic SVM algorithm from mathematical point of view is shown below:

1.Let's assume that one has input and output datasets X and Y, where x ∈ X, y ∈ Y and the training set (x_1,y_1)………(x_n,y_n)

2. The goal is to learn a classifier y = *f* (x,α), where α are the parameters of the Function

3. Function *f* (x, α) learns by choosing a function that has minimum error rate, which is calculated by

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i,\alpha),y_i) \quad \text{-------(1)}$$

where *l* is zero-one loss-function *l* $(y,y^1) = 1$, if $y \neq y^1$ and 0 otherwise. R_{emp} is al- so called the empirical risk

4. This helps us to minimize the overall risk

--------(2)

$$R(\alpha) = \int l(f(x,\alpha),y)\,dP(x,y)$$

Where *P*(x, y) is unknown joint distribution function of x, and y

5. Finally, we choose the set of hyperplanes, so

$$\frac{1}{n}\sum_{i=1}^{n} l(f(w \cdot x_i + b, y_i) + ||w||^2 \quad \text{-----(3)}$$

Is a subject to min_i " = |w .x_i| 1, where b is bias unit and || *w* $||^2$ is a complexity term and can be optimized in different ways depending on the actual task.

LOGISTIC REGRESSION (LR):

The Logistic Regression is one of the methods used in binary classification which aims at predicting an instance member ship to any of two class labels. Although logistic regression may sound as if it is used for the regression instead of the classification, the former is largely used for the latter one. The algorithm simulates the relationship between the independent variables (features) and the logistic function, commonly referred to as the sigmoid function, is that, given a higher value of sensitivity, the probability of an outcome being classified into a particular class increased whereas a low value of specificity reduces the possibility of an outcome belonging to a specific class. The sigmoid function is one of the reasons that makes the prediction result to lie between the bounds of 0 and 1 in this way it is possible to interpret the generated output as a probability of being in a specific class.

Maximum likelihood estimation is a technique used for the estimation of the model parameters in parameters in logistic regression so that the maximum likelihood of the given set of outcomes is obtained. Ultimately, the model yields a decision boundary that lies between both sets of the class in this feature space. Since Logistic Regression exhibits these desirable features, this method is much popular in many disciplines that include the fields of medicine and finance in particular, and marketing as well.

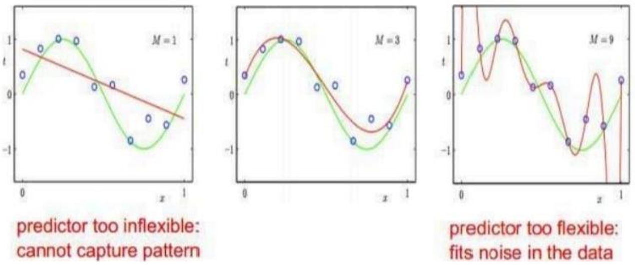


Fig:4 Regression

The logistic regression hypothesis is based on the logistic function (sigmoid function) and is expressed as:

$h(x)=\sigma(b_0+b_1 \cdot \text{feature}_1+b_2 \cdot \text{feature}_2+\ldots+b_n \cdot \text{feature}_n)$
where:

$h(x)$ is the predicted probability that the water sample belongs to the positive class (e.g., contaminated),

$$\frac{1}{1+e^{-z}} \quad \frac{1}{1+e^{-z}}$$

σ(z) is the logistic (sigmoid) function: ,z is the linear combination of the coefficients(b_0,b_1,..,b_n) and features(feature_1,feature_2,…,feature_n):z= b_0 + b_1.feature_1 + b_2.feature_2 +...b_n. feature_n

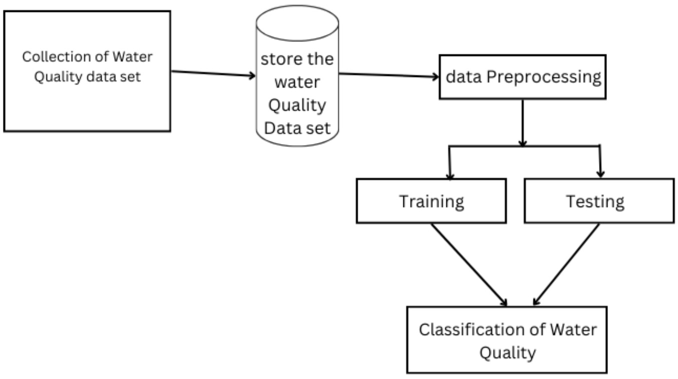

Collection of Water Quality data set
store the water Quality Data set
data Preprocessing
Training
Testing
Classification of Water Quality

Fig: 5 System Architecture

We have Collected the data from the various websites. That dataset having Hours, Day, Monthly values we can extract that dataset for further analysis.

The collected datasets are stored in the form of folder in your system.

Training and Testing : In training we will train the dataset with various values and then we will go to testing part. In this testing we will use different algorithms to check weather water is contaminated or Drinking water.

Classification : Finally In classification by using SVM and Logistic Regression algorithms we will classify our data in linearly and non-linearly and We will create a module to Predict the water Leakage and then it will provide the appropriate results.

MODULES DESCRIPTION

CREATE A WEB INTERFACE:

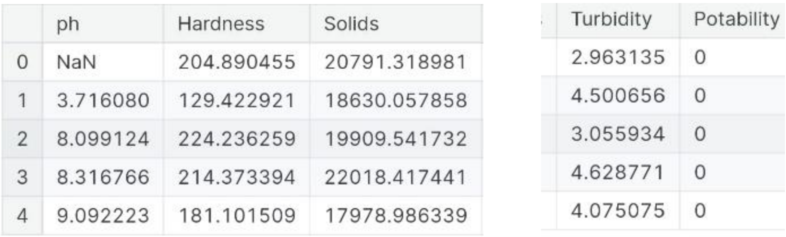
Web interface design is a process of constructing an engaging user interactive platform that allows website or web applications users to interact with it naturally. The UX layout should emphasize simplicity and easy flow for users, with a user-friendly navigational design that makes moving through the site's functions seem nothing less than magical. Along with it, relevant elements like menus, buttons, and forms are to be placed deliberately so that the information provided is rationally structured and logical activities are taken in sequence. Visual element aspects, for instance color themes and typography play out an auxiliary role in the formation of a general user- experience that also contributes to the idea of having consistent interface.

In order to meet the needs of desktop, tablet and smartphone users a responsive web design is required since it enables provisioning of consistent user experience enabling a much better enjoyment in browsing. As a norm, structuring the content using HTML, styling them with CSS and interacting utilizing specifically JavaScript are phycologically commonplace things. One of the crucial features that web interfaces usually entail can be described as dynamic ones – asynchronous data loading, real-time updates and enhanced users' involvement. Consistent testing and user input are instrumental in polishing up its web interface to fit user assumptions and avoid possible drawbacks associated with usability such as committing mistakes.

DATA COLLECTION:

Data collection is an integral part of the process that involves capturing information with the aim of analysing it andmaking a choice. Respondent It requires a systematic collection of empirical or factual material pertaining to aparticular study question or problem. This process greatly influences the choice of data sources and methods, which appears to be heavily based on the subject matter in need of investigation. Data sources can either be through surveys, or interviews, sensors,, experiments and or scraping online platforms. The design of data collection, which depends on the accuracy and quality of gathered information is the best life way in ensuring validity, reliability of any given analysis as well as analysis that derive from this information.

Researchers and organizations are ethically bound to respect sacredness of privacy and confidentiality coveted by informants, ought to select guidelines that minimise respondent bias, committee norms that ensure inhabitant anonymity with individual information safeguarded as indicated by her wishes. Moreover, the amount and kinds of collected data have increased significantly due to technological progress, leading to a phenomenal rise in big data problematics. The currents changes are efficient working with large sets of data, the security of data and convenience while biasing in the collection processes are actual problems arising from an ever-changing nature of science and research.



	ph	Hardness	Solids	Turbidity	Potability
0	NaN	204.890455	20791.318981	2.963135	0
1	3.716080	129.422921	18630.057858	4.500656	0
2	8.099124	224.236259	19909.541732	3.055934	0
3	8.316766	214.373394	22018.417441	4.628771	0
4	9.092223	181.101509	17978.986339	4.075075	0

Table 3: Data collection

DATA PREPROCESSING:

Raw data, which has been cleaned and prepared such that it can be ready for analysis or machine learning is known as the process of preprocessing data. It pertains to cleaning, transformation, and refining data from both quality as well usability aspect. Handling missing values, removing duplicate values, outlier adjustments are the standard techniques. Standardization and normalization make certain data sets have consistent scales. The encoded numeric encoding may be categorical column variables that are compatible of model. Feature scaling maps numeric values into a compatible range. Imputation methods are based on patterns or statistical measures and substitute heterogeneous missing values. Some processes text data can be tokenized and stemmed. Normalization of data guarantees that all data has to adopt the desired format. Preprocessing is a repetitive procedure, that consists of exploration, transformation and confirmation stages; But the aim is to improve data quality, creating it as a desired state for analysis or model training.

TRAINING MODEL:

Machine learning models are trained on the training data by giving the algorithm instructions to identify patterns and use them in predicting outcomes from more input data. In most cases, the process kicks off by feeding a labeled dataset to the model; over time, the algorithm learns how I tie in input features and corresponding output labels. In training, the model refines its internal parameters through iterative optimization and minimizes the gap between the desired output of"1″ or "-1", depending on whether it was correct or incorrect with actual outcomes. Typically, this optimization depends on the loss function that measures some performance of the model. The process of training is iterative. In editing regarding the model's generalization, there are bypasses and backward passes, The duration and success of the training process is a function the complexity of the model, and quality of data from which it is trained, as well as optimization techniques used.

Model training requires a balancing act, aimed at ensuring that it is neither overfitting – where the model becomes too specialized towards the training data to generalize well or performs an underfit which tends to oversimplify and miss cues for patterns. Tweaking of hyperparameters is an important parameter for gaining high accuracy and optimizing the model's performance, and hyperparameter tuning adjusts settings that include learning rates or regularization parameters. Validating the performance of trained model is done using another seperate validation dataset to check whether it has generalization capability. After being comfortable with the results, it can be employed to make predictions on new data that was not used during training, demonstrating where the train cycle comes to an end and the model has learned patterns to develop knowledge how to act.

TESTING MODEL:

In machine learning, test module often stands for a body of data or a specific step in the process of model testing.The process of testing involves evaluation of a trained model on another dataset that can be used to determine the actual performance of the model.This allows to us to determine the prediction performance of the model when it is applied to new, data set.

SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine (SVM) is one of the most important algorithms which are commonly used as a baseline to compare model performances during the research." It is a linear model-based family, where training consists of transforming the original vector into a more dimensions space to finding a dividing hyperplane that would have the highest distance

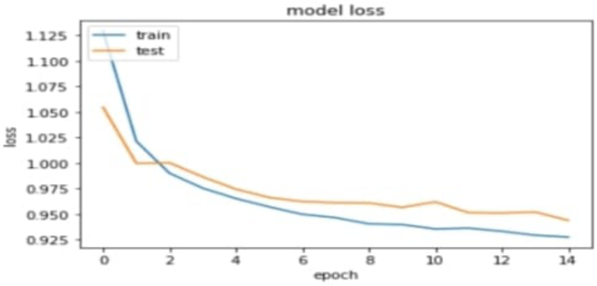


Fig6: Accuracy of model in SVM

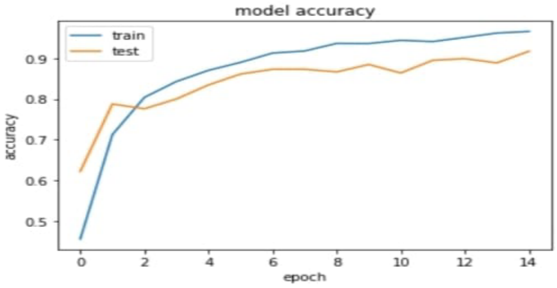


Fig7: Accuracy Loss of model

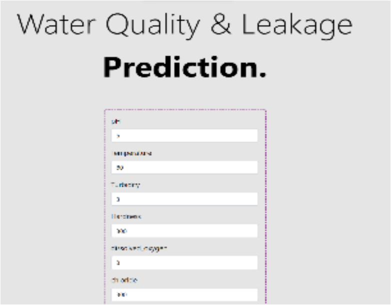
IV. PREDICTING RESULTS

The forecasting of results by utilizing a trained machine learning model includes when the patterns that were learnt are used on fresh, unseen data to make projections or classifications. After being trained and validated, the model can then be utilized using a field scenario system to predict real-world situations. The internal parameters adjusted earlier with the training phase are utilized and predictions or classifications are derived using these inputs. In generalization, the estimates depend more on how intrinsic the model is and its relevance to making accurate predictions from training data set towards unseen observations.

It is important that model accuracy only should be tested along with the new sample of data. This can mean keeping tabs on the metrics of performance, adjusting the hyperparameters, or even retraining with new datasets in order ensure its relevance in changing scenarios. The ability to predict the outcome with high accuracy, reliability, and temporal validity, as well as the model's capacity for stochastic development enable it such power to disclose valuable insights that aid informed decision making in multiple fields from funding to health care systems, image recognition of natural language processing

V. RESULT

The two major factors that influence water quality modeling and prediction are the safety of, and reliability on water resources. In this work, we use the machine learning algorithms to deal with such an issue. The results of this research present machine learning techniques as a promising tool for the simulation of water level and quality, which provides high accuracy. We seek to offer an ideal tool for the evaluation of water quality. With the help of WQI, we can guess water leakage and whether it is contaminated or not fit for drinking.



Water Quality & Leakage

Prediction.

pH
7

temperature
30

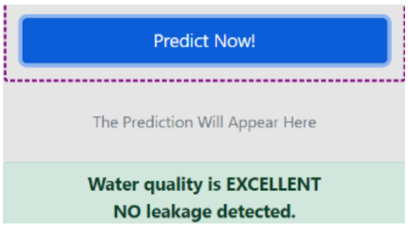
Turbidity
3

Hardness
300

dissolved_oxygen
3

chloride
300

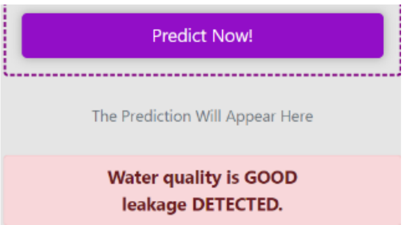
Fig8a: Prediction of water quality



The Prediction Will Appear Here

Water quality is EXCELLENT
NO leakage detected.

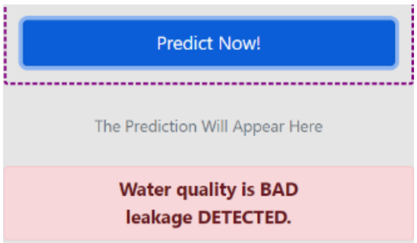
Fig8a: Prediction of water quality



The Prediction Will Appear Here

Water quality is GOOD
leakage DETECTED.

Fig 8c:Prediction of water quality



The Prediction Will Appear Here

Water quality is BAD
leakage DETECTED.

Fig 8d Prediction of water quality

VI. CONCLUSION

This research aimed to assess the effectiveness of machine learning algorithms in predicting the quality of river water and classifying the water quality index. Various machine learning techniques, such as Support Vector Machines (SVM) and Logistic Regression, were utilized to construct WQI classifiers. Data on parameters such as pH, temperature, turbidity, hardness, dissolved oxygen, chloride, and pressure were collected, modeled, and incorporated into the models. The performance of the river water quality index was evaluated using these models. Additionally, the water quality index was analyzed using SVM classifiers and Logistic Regression classifiers, along with other models. Furthermore,the use of hybrid models incorporating deep learning algorithms was explored as a means to enhance the efficiency of water quality prediction.

REFERENCES

[1] C.V.Sillberg,P. Kullavanijaya, OChavalparit, Water quality classification by integration of attributerealizationandsupportvectormachineforthechaophrayariver,JournalofEcological Engineering 22 (2021)

[2] M.Yilma,Z.Kiflie,A.Windsperger,N.Gessese,Applicationofarti-ficialneuralnetworkinwater qualityindexprediction:acasestudyin littleAkaki River,AddisAbaba,Ethiopia,ModelingEarth Systems and Environment 4 (2018)

[3] Y.R.Ding,Y.J.Cai,P.D.Sun,B.Chen,Theuseofcombinedneuralnetworksandgeneticalgorithmsfor prediction of river water quality, Journal ofApplied Research and Technology 12 (2014)

[4] U.Ahmed,R.Mumtaz,H.Anwar,A.A.Shah,R.Irfan,J.García-Nieto,Efficientwaterquality prediction using supervised machine learning, Water 11 (2019),

[5] Zhang,J.,Zhu,X.,Yue,Y.,&Wong,P.W.Areal-timeanomalydetectionalgorithm/orwater quality data using dual time-moving windows. 2017 Seventh international conference on innovative computing technology (INTECH) (pp. 36–41). IEEE, (2017).

[6] Sakizadeh,M.Artificialintelligenceforthepredictionofwaterquality indexin groundwatersystems. Model. Earth Syst. Environ. (2016)

[7] Fitore Muharemi, Doina Logofătu& Florin Leon (2019) Machine learning approaches for anomaly detectionofwaterqualityonareal-worlddataset,JournalofInformationandTelecommunication,3:3, 294-307(2019)

[8] TejasSubramanya,DavitHarutyunyan,RobertoRiggio,Machinelearning-drivenservicefunctionchain placement and scaling in MEC-enabled5G networks, Computer Networks, Volume 166, 2020,106980, ISSN 1389-1286

[9] J.P.NairandM.S.Vijaya,"PredictiveModelsforRiverWaterQualityusingMachineLearningand BigDataTechniques-ASurvey,"2021InternationalConferenceonArtificialIntelligenceandSmart Systems (ICAIS), (2021)

[10] Kalimur Rahman, Saurav Barua, H.M. Imran, Assessment of water quality and apportionment of pollutionsourcesofanurbanlakeusingmultivariatestatisticalanalysis,CleanerEngineeringand Technology, Volume 5, 2021, 100309, ISSN 2666-7908(2021)

[11] Arunkumar,R&Thambusamy,Velmurugan.(2021).AnExploratoryDataAnalysisProcesson Groundwater Quality Data. 54. (2021)

[12] MarisolVega,RafaelPardo,EnriqueBarrado,LuisDebán,Assessmentofseasonalandpollutingeffects on the quality of river water by exploratory data analysis, Water Research, Volume 32, Issue 12, 1998, Pages 3581-3592, ISSN 0043-1354

[13] StroombergG.J.,FreriksI.L.,SmedesF.andCo®noW.P.(2020)InQualityAssurancein Environmental Monitoring, ed. P. Quevauviller. VCH, Weinheim.(2020)

[14] GTan,JYan,CGao,andSYang,Predictionofwaterqualitytimeseriesdatabasedonleastsquares support vector machine, Procedia Engineering, Vol. 31, (2012)

[15] WCLeong,ABahadori,JZhang,andZAhmad,Predictionofwaterqualityindex(WQI)usingsupport vector machine (SVM) and least square-support vector machine (LS-SVM), International Journal of River Basin Management, Vol. 19, (2021)