MONITORING AND ANALYSING WATER QUALITY USING WIRELESSSYSTEM AND MACHINE LEARNING

[1]LAVANYA SINGH, [2]SHAIK ABDUL KALAM, [3]DR SIVAKUMAR S

[1,2]BTECH STUDENT AT VELLORE INSITUTE OF TECHNOLOGY, CHENNAI
[3]ASSOCIATE PROFESSOR GRADE 2, SENSE VELLORE INSTITUTE OF TECHNOLOGY, CHENNAI

Abstract- **Water is the most essential need for all forms of life. Predicting water quality is of prime importance while formulating the environmental control plan and can contribute to better water resource conservation. Accurate projections of water quality are the testimony that can assist authorities in making prudent choices before predicament. The goal of this study was to provide a novel Machine Learning based model for water quality prediction and it aimed for comparative analysis of different Machine Learningalgorithms on the available dataset while evaluating their accuracy. Intensive and comprehensive approach was performed to study the potability of water. A software simulation using three sensors namely temperature, turbidity and pH sensor on Proteus platform was implemented initially. Then in the second stage of the study, a hardware project using these 3 sensors was set up with Arduino and ESP8266 to test real-time various water implies and store these values in cloud server using Node MCU and predict their potability. An extensive study was carried out to explore further the categorisation of potable water using various Machine Learning algorithms in the third and final stage of the current project. Balancing the data set using re-sampling and shuffling helped to improve the accuracy of the models and prevent bias towards the majority class. The Random Forest algorithm (RF) was executed the best with an accuracy of 88 percent.The Decision Tree and XGBoost algorithms also performed well, achieving accuracies of 80 percentand 86 percent respectively. The SVM and ANN algorithms performed inferiorly, achieving accuraciesof 70 percent and 68 percent respectively. The KNN and AdaBoost algorithm under performed with 66 percent and 63 percent accuracies respectively Performance of the ML techniques were also evaluated usingaccuracy precision, recall, F1 Score and MCC score which reconfirmed the highest performance of RF algorithm. Performance of all the ML algorithms were compared against Deep Learning(ANN), it was foundall the tree based classifiers(ML algorithms) outperformed the Deep Learning algorithm (ANN).with RF showing the best accuracy of 88 percent. It can also be reasonably concluded and deduced that deep learning algorithms have limited performance on tabular and numerical data which is not linearly separable. DL algorithms are superior for images and text. These results suggest that the RF algorithm isa promising approach for classifying potable water and can be used for future research in this area. Asa future scopethe IOT based hardware system of this present study can be explored to testing real-time water samples andanalysing their potability.**

Index Terms: **Random Forest, ANN, XGBoost, KNN, Adaboost, Decision Tree and SVM**

I. INTRODUCTION

Water is vital for all forms of life that exist on Earth. Water requirement for humans differs with age, gender and place ofliving. An adult male need 3 litres/day whereas female needs

2.3 litres/day. The exceptional characteristics of water make it basic to life and its excellent ability to dissolve various substances allow our cells to perform various biochemical reactions and use these valuable ingredients. Water can be grouped into ground water and surface water depending on itssource[1]. Anthropogenic activities have put the quality

VOLUME,

of water at risk that people consume which may includepollutants such as toxic and hazardous waste, pesticides, fertilizers and industrial effluents. Water may be classified [1]as drinkable, pleasant, contaminated (polluted), and infected depending on its quality. Water that is fit for human con- sumption, has a good flavour, and is potable. Consideration is given to the presence of compounds in water that are both visually pleasant and do not provide a health risk. Contaminated (polluted) water possess undesirable biolog- ical, physical, chemical or radiologically active substances, making it unfit for drinking or household usage. Infected water harbours pathogenic organisms[1]. Water quality pa- rameters determine the chemical, biological and physical properties of water. Predicting water quality is of prime importance while formulating the environmental control planand can contribute to better water resource conservation. Or- ganisations for Water management have installed monitoringstations to determine and track the development of the water quality issues. Accurate projections of quality of water are thetestimony that can assist authorities in making prudent choices before predicament.
In this study our objective was to provide a novel MachineLearning-based model for predicting water quality and aimedcomparative analysis of different Machine Learning algo-rithms on the available dataset and evaluate their accuracy.

A. *ABBREVIATIONS*

1) ANN- Artificial Neural Network
2) HTML-Hyper Text markup Language

3) FP- False Positive
4) FN- False Negative
5) KNN- K-nearest neighbors
6) MCC - Matthew’s correlation coefficient
7) RF – Random Forest
8) SVM- Support Vector Machine
9) TP-True Positive
10) TN- True Negative

B. RELATED WORKS

Through this research paper, evaluation of different methods employed to analyse the water quality available to us was done by using various Machine Learning techniques and alsopredicted the water quality which shall be available in the nearfuture. The present study investigated and comparedthe accuracy achieved by the different techniques and deter- mined which technique predicts most efficiently. Artificial Neural Network (Deep Learning technique) was also imple- mented to further analyse the water quality for precision and accuracy of prediction of the water quality. Manisha Korangaet al [2] evaluated water quality for Nainital Lake. In this study, regression analysis was employed on eight ML algo- rithms and classification analysis on nine ML algorithms. Random Forest algorithm was the most proficient model to forecast the water quality of Nainital Lake using regression analysis. The accuracy, precision, recall, and F1 score of Random Forest and Support Vector Machine were superior, at 0.98742, 0.98799, and 0.98742 respectively. In terms of classification methods, no one technique was deemed to be the best; Random Forest, Stochastic Gradient Descent,and Support Vector Machine all performed precisely and accurately.

Water Quality Index (WQI) calculations using MachineLearning (ML) were given by Sandeep Bansal and G.Geetha

[3] and categorised water quality to predict water qualities foruse. When classifying water quality using the Decision tree technique, 98.28 percent accuracy was attained.The average error was determined to be 0.0199, while the root mean squareerror was found to be 0.0996. In this article, theWEKA tool(a Java platform) was used to perform q value normalisation.Deep learning (DL) based models for assessing ground- water quality were introduced by Sudhakar Singha et al.

[4] and compared to three different Machine Learning (ML) models: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). Due to its ability to examine complexity and nonlinear interactions ina dataset, the DL approach became more important. Theprimary factor in this ML technique’s success was that it disregarded the specific feature criteria that, when compared to conventional ML methods, were the most representative. The sole drawback of this study was that the predictionmodelsonly took into account one monsoon dataset. The authors of this article employed a feed forward neural network and the EWQI.

In their study, Liang Kuangi et al. [5] introduced the KIG- ELM hybrid DO prediction model, which included K-means,extreme learning machine (ELM), and improved genetic algorithm (IGA). The combination of K-means, IGA, and ELM was successfully used to address the issue of low accuracy of a single model. More than 90 percent of predictions made using the six models were accurate.

S.Angel Vergina [6] used a Real Time Water Quality Monitoring using ML algorithm sensor for measuring the totalamount of dissolved solvents and hydrogen ions inthe water. K Means calculation was done to forecast the character of water. For a better understanding of water quality affordable embedded devices like the Raspberry Pi and Arduino Uno were used. This proposed concept provides rural residents with water of a consistent quality. Since water samples were taken from different sources (Mud water, Lemon water, Saltwater, Tap water, and Drinking water) k means clusteringhas given the efficient solutions. In this paper no pre- processing was done for the collected data.

Hamza Khurshid et al [7] investigated bacterial prediction using internet of things (IoT) and Machine Learning. The sensor nodes, which are placed around the study area atvarious locations, used GSM modules for sending, process- ing and analysis of data. The data gathered over severalmonths was utilised to classify the quality of water using water quality indicators and to anticipate the presence of microorganisms using algorithms of Machine Learning.For data visualisation, a Web portal with a dashboard of Web services was developed to display the heat maps and other related info-graphics. IoT nodes were used to gather real- timeinformation on water quality. The Rawal Lake FiltrationPlantprovided historical data. For the prediction of faecal coliformbacteria, many ML algorithms and neural networks (NN), (CNN), ridge regression (RR), (SVM), (DTR), and (BR) weretrained. The highest performance was demon- strated by SVMand Bayesian regression models, whose mean squared errors (MSE) were respectively 0.35575 and 0.39566. The few benefits listed are remote monitoring, scalability, real-time water quality monitoring, and portable gear. In this research data collected by IoT nodes was a pointdata and the authors used GSM module for transferring the data.

Mourade Azrour1 et al [8] for prediction of efficient waterquality analysed various Machine learning algorithms. The four water characteristics of temperature, pH, turbidity, and coliforms were the foundation of their strategy. In order to estimate the water quality index, several regression techniques had proven to be useful and effective. Addition- ally, the use of artificial neural networks offered the most effective method of categorising water quality and to developa capable model for estimating the water quality class and index.

P. Kirankumar et al [9] used IOT and ML to study Smart Monitoring and Water Quality Management in Agriculture aswell as marine water and underground water. Implementationof the suggested framework using an Arduino and Machine Learning was done. The Arduino Uno uses sensors and fac- tors that are anticipated in advance using Machine Learning techniques to collect data on things like pH, temperature, dissolved oxygen, and turbidity in water. The farmer will receive a warning notification if the desired range of the parameters was exceeded so that he can make the necessary changes. By using the Machine Learning algorithms, it can predict the most accurate values and send the notificationto the farmer when the range of water parameters exceeds or decreases the ranges. This helps the farmers to know the values automatically and take the necessary action.

II. IMPLEMENTATION

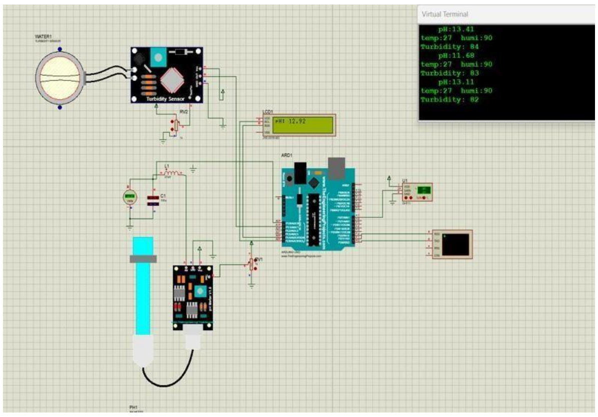
A. *SOFTWARE SETUP*

WORKING OF SOFTWARE MODEL

1) The LC Circuit was connected to the pH sensor and to the Arduino and the potentiometer was connected to thetest pin. Next from the library DHT11 temperaturesensor was imported.

2) Temperature and turbidity sensors were connected to the port of the Arduino uno, Terminal display is added.

3) Test pin was connected to the variable resistance. Fromthe LCD display connections SCL and SDA pins were also attached to the Arduino.



```
    pH:13.41
temp:27  humi:90
Turbidity: 84
    pH:11.68
temp:27  humi:90
Turbidity: 83
    pH:13.11
temp:27  humi:90
Turbidity: 82
```

FIGURE 1. Proteus Simulation

4) Code for virtual terminal was set up. Code loop for running the sensors was added and 6 values were takenfrom the sensors and their average was calculated, which was taken as the reading of pH, temperature and turbidity (Fig 1).

B. *HARDWARE SETUP*

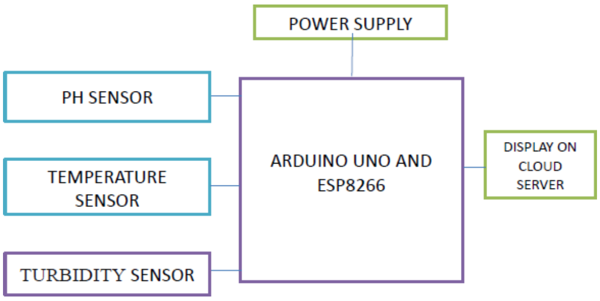


FIGURE 2. BLOCK DIAGRAM OF HARDWARE

COMPONENTS OF THE HARDWARE

1) PH SENSOR
2) TEMPERATURE SENSOR
3) TURBIDITY SENSOR
4) ARDUINO UNO
5) ESP8266(WIFI MODEL)
6) BREADBOARD
7) POWER

WORKING OF HARDWARE SETUP(Fig 6)

ARDUINO UNO: The temperature sensor(Fig 5) usesa one-wire bus for communication and was connected tothe Arduino's digital pin(ONEWIREBUS). The turbidity sensor(Fig 4) and pH sensor(Fig 3) were attached to theanalog pin on the Arduino. The Node MCU was affixed to theArduino's digital receiver and transmission pins using a software serial communication. ESP8266: To establish serialcommunication between the Arduino and the Node MCU, the serial pins were connected together. The Node MCU used software serial library which were connected two digital

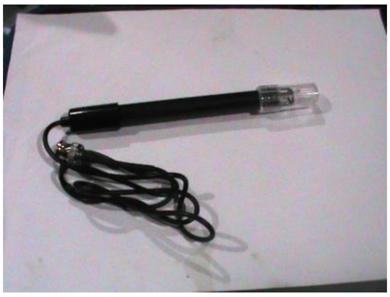


FIGURE 3. PH SENSOR

pins of the Arduino to the serial pins of the Node MCU. The connections were made in such a way that the transmission ofone device is connected to the receiver of the other and vice versa so that the data is transmitted from the Transmission pinof one device to the Receiver pin of the other. Arduino

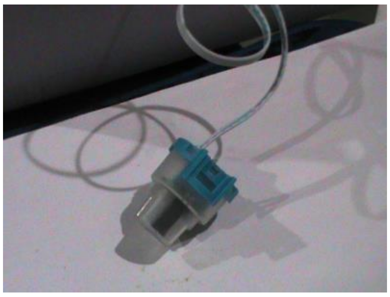


FIGURE 4. TURBIDITY SENSOR

code collects data from three sensors, a temperature, pHand turbidity sensors, and sends the data to a NodeMCU using theSoftware Serial library. The code uses the One Wire library to communicate with the temperature sensor,the Dallas Temperature library to obtain the temperaturereadings, and the Arduino Json library to format the data asa JSON object before sending it to the Node MCU. The codefirst initializes the serial communication with the Node MCUand the serial monitor with a baud rate of 9600. In the loop() function, the readings from the three sensors are collected and stored in a JSON object. The data is then sent to the

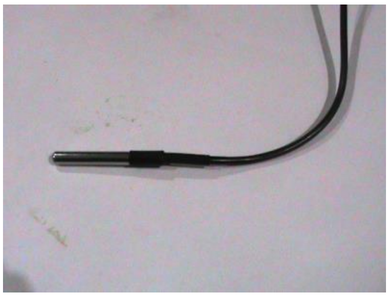


FIGURE 5. TEMPERATURE SENSOR

NodeMCU using the printTo() method. The temp probe() function reads the temperature in Celsius and converts it to Fahrenheit. The turbidity() function reads the value fromthe turbidity sensor, and the pH() function reads the pH value from the pH sensor. Finally, the code has a shortdelay of 2000 milliseconds between each reading to avoid overloading the Node MCU with too much data at once. NodeMCU (ESP8266): Code sends data from an Arduinoto aNode MCU via serial communication. The code setsup a web server on the Node MCU, which displays thedata from the sensors affixed to the Arduino. The data from the Arduinois sent to the Node MCU using the Software Serial library. The code starts by including the necessary libraries for serial communication, Wi-Fi connectivity, and server handling. Thenetwork credentials, such as the Wi-Fi name (SSID) andpassword, are then defined. An instance of the ESP8266 WebServer class is created and defined to listenon port 80. In the setup function, the serial communication with both the Node MCU and Arduino is initialized. The

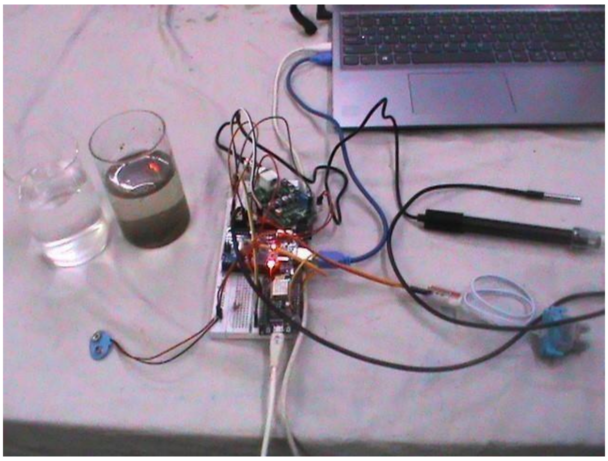


FIGURE 6. IMAGE OF WORKING HARDWARE SETUP

Node MCU then connects to the Wi-Fi network using the defined credentials. The web server is started and a simple HTML page is created, which displays the data from the sensors in the manner of an inner HTML text. In the loop function, the code uses the Static Json Buffer class from the Arduino Json library to parse the incoming serial data from the Arduino, which is expected to be in JSON format. The values for the sensors, such as temperature and turbidity, are extracted from the JSON object and saved in variables. Thesevariables are then displayed on the HTML page of the web server.

C. MACHINE LEARNING AND DEEP LEARNING

A subset of artificial intelligence called Machine Learning updates its knowledge every time new data is presented and produces outcomes based on previously learned information.It takes a dataset as input that contains prior data about the task at hand, constructs a model utilizing the information using algorithms, and then can provide results for the input that isn't available in the dataset by providing the closest or most accurate prediction. Machine Learning models help in data representation, evaluation and optimisation of models.

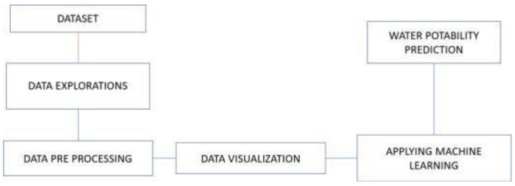


FIGURE 7. PROPOSED DIAGRAM OF MACHINE LEARNING

Machine Learning classifications algorithms were used topredict potability of water.

1) Support Vector Machine
2) Adaboost
3) RF(Random Forest)
4) XGBoost
5) Decision Tree
6) ANN(deep learning)
7) KNN

Steps undertaken while Machine Learning included Im- porting necessary libraries, reading and analysing data, per- forming deep exploratory data analysis, performing data visualization, data pre-processing, splitting data, scaling the data building models and predicting the accuracy using algorithms of Machine Learning.

The dataset was taken from open source Kaggle, using displaying head to display top 5 results and bottom 5 results was done. Data pre-processing was performed and libraries were imported. Null values in the dataset were checked,targetvariable distribution was displayed, bar plot followed bydropping of null values was created, values were furtherreduced and bar plots were created.(Fig 12,13)

To balance the dataset and to increase the model accuracy, resampling was done and shuffled to fill null values(Fig 12). After the dataset was balanced, the bar plot was analysed using the The linear link between two variables is measured by the Pearson correlation coefficient (Fig. 13).

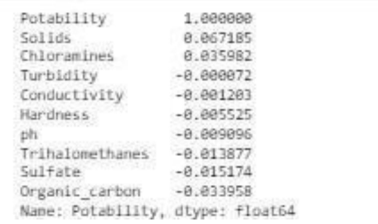
All the features' pair plots and the co-relation between them were plotted. Dataset standardisation was carried out. Data modelling was the following phase. The dataset was divided into two sections. The training dataset was the first subset thatwas used to fit the model. The second subset was not used to train the model; instead, the input element from the dataset was used to produce predictions, which were then compared to actual values. It's the second dataset, the test dataset.

To assess and study a classification model’s efficacy, a confusion matrix of size N x N is utilised, where N is the totalnumber of target classes. The matrix contrasts the actualgoal values with projected values from the Machine Learningmodel. This gave a comprehensive picture of the catego- rization models’ abilities and the types of mistakes they werecommitting. Hyperparameter tuning was performed for model optimization using hyperparameter tuning for Decision tree, XGBoost and RF algorithm.

III. RESULTS AND DISCUSSSION

An open-source dataset of 3277 rows which was collected from Kaggle website and Machine Learning was performed on following 10 parameters namely:- pH, Hardness, solids, chloramines, sulfate, conductivity, Trihalomethanes, Turbid- ity and Potability. Using Google Colab with python pro- gramming language and Database from the open source Kaggle and Random Forest classifier along with other clas- sification models such as Decision tree classifier, Support Vector Machine, XGBoost, Adaboost,KNN and ANN ar-tificial neural network(deep learning) were implemented. Data pre-processing was carried out by importing libraries, Extraction of independent variable, x-axis and y -axis was done followed by data optimization and visualization. The dataset includes 3276 entries ranging from 0 to 3275. Null values, non- null values and float values were checked. Null values were dropped and zero was given if the water sample was non potable and one for Potable water which has been represented on a bar plot, corresponding number of entries: 0= 1998 and 1= 1278 (Fig 11, 12). Balancing of data implies fitting the Random Forest, Decision tree, XGBoost and otheralgorithms to the training set. As Data imbalance hides the true performance of the model which is objectively not good therefore dataset was balanced (Fig 13).

Balancing of dataset was done so that number of potable and non-potable entries were equal, inorder to analyse the execution of the various algorithms (Fig 13). Pearson’s Coefficient of Correlation was performed for better data vi- sualisation by plotting correlation between features and pair plots for all the features were prepared (Fig 15). Pearson’s Correlation was carried out and Correlation Matrix was laid. Corr mat graph, heatmap (Fig 14) and table showed max- imum value with solid 0.067185 followed by chloramines 0.035982, turbidity -0.000072, pH -0.009096 and minimum correlation with organic carbon was -0.033958(Fig 8).



```
Potability         1.000000
Solids             0.067185
Chloramines        0.035982
Turbidity         -0.000072
Conductivity      -0.001203
Hardness          -0.005525
ph                -0.009096
Trihalomethanes   -0.013877
Sulfate           -0.015174
Organic_carbon    -0.033958
Name: Potability, dtype: float64
```

FIGURE 8. CORRELATION VALUES

Splitting the dataset into training and test set was per- formed and all the algorithm namely Adaboost Classifier, Decision Tree Classifier, Random Forest Classifier, XGBoostclassifier were imported. The independent and dependentvariables were extracted and standard scaling was performed.All the algorithms were run and the test set result werepredicted. Since the test model was fitted to the training set, so the results could be predicted. For prediction, a new prediction vector was created. Hence, the confusion matrix could be printed. Hyperparameter tuning was performed for model optimization using hyperparameter tuning librariesfor XGBoost, Decision tree and Random Forest algorithms. Incase of KNN classifier, an error versus k values graph was plotted. Deep Learning algorithm ANN was also modelled and its accuracy was also optimized using optimiser hyper- parameter tuning.

A. *OUTPUT ANALYSIS*

HARDWARE OUTPUT

Using the hardware which was developed using 3 sensors, Arduino UNO and ESP8266 a few real-time water samples were tested and results were stored in the cloud server. The water samples were analysed for their pH, turbidity and temperature values (Fig 9)

B. *OUTPUT OF WATER TESTING*

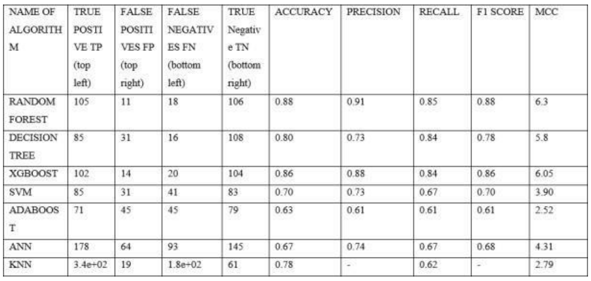
	SAMPLE 1	SAMPLE 2
PH VALUES	6.14	10.59
TURBIDITY VALUES(NTU)	402.00	546.00
TEMPERATURE (IN Fahrenheit)	71.04	73.17

FIGURE 9. SENSOR TO NODE MCU WEB SERVER

C. *MACHINE LEARNING OUTPUT*

Creating the confusion Matrix :

The confusion Matrices were created to determine and predictthe accuracy of the algorithm using the codes.



NAME OF ALGORITHM	TRUE POSITIVE TP (top left)	FALSE POSITIVES FP (top right)	FALSE NEGATIVES FN (bottom left)	TRUE Negative TN (bottom right)	ACCURACY	PRECISION	RECALL	F1 SCORE	MCC
RANDOM FOREST	105	11	18	106	0.88	0.91	0.85	0.88	6.3
DECISION TREE	85	31	16	108	0.80	0.73	0.84	0.78	5.8
XGBOOST	102	14	20	104	0.86	0.88	0.84	0.86	6.05
SVM	85	31	41	83	0.70	0.73	0.67	0.70	3.90
ADABOOST	71	45	45	79	0.63	0.61	0.61	0.61	2.52
ANN	178	64	93	145	0.67	0.74	0.67	0.68	4.31
KNN	3.4e+02	19	1.8e+02	61	0.78	-	0.62	-	2.79

FIGURE 10. TABLE OF RESULTS

After balancing the dataset and identifying the important input variables, several Machine Learning algorithms were used to classify potable water. The outcome of the analysis showed that the Random Forest algorithm performed thebest,achieving an accuracy of 88 percent on the testingset (Fig 18). The XGBoost and Decision Tree algorithms followed closely, achieving accuracies of 86 percent and80 percent (Fig 16,17), respectively. The SVM and ANN algorithms performed inferiorly, achieving accuracies of 70 percent and 68 percent, respectively (Fig 19,21). The Ad- aBoost algorithm also achieved a low accuracy of 63 percent (Fig 10,20).

1) Accuracy= TP+TN/TP+TN+FP+FN
2) MCC (Matthew's correlation coefficient) =TP*TN-FP*FN/(TP+FP)*(TP+FN)*(TN+FP) *(TN+FN)
3) PRECISION=TP/TP+FP
4) RECALL=TP/TP+FN
5) f1 score=2*precision*recall/(precision+recall)

A confusion matrix was built to evaluate the performance of each algorithm. The confusion matrix showed that the Random Forest algorithm had the highest number of true positives and fewest false positives and false negatives(Fig 10).

Evaluating the scores of true positive,true negative,false pos-itive and false negatives, accuracy, precision, recall and F1 scores were calculated.The F1 scores are classically between 0 to 1. High F1 scores show high accuracy of the model. TheF1 scores for tree based classifiers namely RF,Decision Tree and XGBoost are high 0.88,0.78 and 0.86 respectively. RF algorithm had shown the maximum F1 scores and highest accuracy.

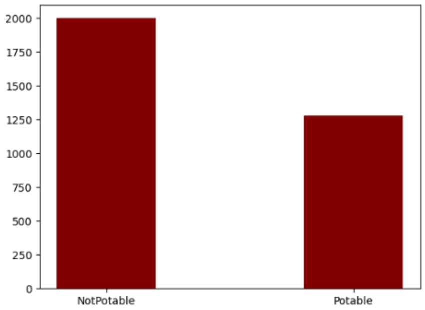


FIGURE 11. BAR GRAPH BETWEEN POTABLE AND NON POTABLEBEFORE DROPPING NULL VALUES

Using Matthew's correlation coefficient, the effectiveness of the algorithms was further examined.Matthew's coefficient of correlation, or MCC, is a specific kind of Pearson correlationcoefficient used in binary classification situations when the prediction and label are two random variables. Matthew's Correlation Coefficient is a discrete case of Pearson Correla- tion Coefficient, in other words. Maximum MCC score(Fig 10) with Random Forest classifier of 6.3 was found, thus indicating that our model prediction using the test dataset is reliable and there is very less chance of over prediction.

Accuracy of each algorithm by plotting a bar graph was checked(Fig 18). The graph demonstrated how accurately(88percent) the Random Forest method outperformed the other algorithms, followed by the Decision Tree and XGBoostalgorithms,80 percent and 86 percent respectively. The SVMand ANN algorithms had similar accuracy, 70 percent and 68 percent respectively, while the AdaBoost algorithm had the lowest accuracy 63 per cent. KNN classifier is based on

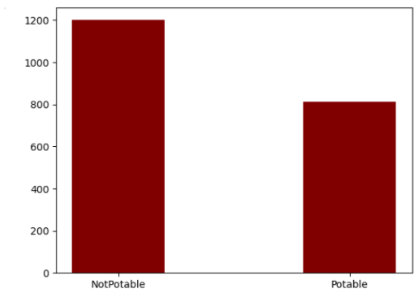


FIGURE 12. BAR GRAPH BETWEEN POTABLE AND NON POTABLE AFTER DROPPING NULL VALUES

k nearest neighbor also showed accuracy of 66 percent (Fig 22) and it was observed in the plot that error decreased asK value increased(Fig 26).

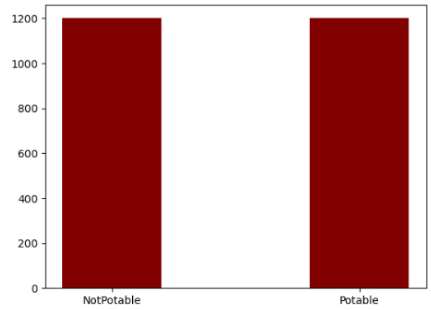


FIGURE 13. BARGRAPH BETWEEN POTABLE AND NON -POTABLE WITH BALANCING

D. NEW APPROACH OF THE MACHINE LEARNINGMODEL

The dataset was taken from open source kaggle.com. A number of previous studies are available with accuracy level of between 59 percent to 63 percent however the ML algorithms used in the present study are unique andgiving accuracy as high as 88 percent. The confusion matrix prediction had been tested on the test set showing the model in the current study is robust and predictable. The MCC Matthew's correlation coefficient score which is a discrete case of Pearson's correlation coefficient, based on false positives(FP) and false negative (FN) values further proves and justifies the robustness of this model as RF algorithm showed minimum number of false positive values (Fig 10). The accuracy is predicted on basis of true positive (TP) and true negative (TN) values which was also observed to be maximum for Random Forest classifier.

Hence, RF outperformed all other ML algorithms based on accuracy,precision,recall, F1 scores and MCC scores.Therefore,it can be inferred that since the data is numerical data and not linearly separable therefore RF, XGBoost followed by Decision Tree, which are tree-based classifiers showed the best accuracy. In the current study, ML

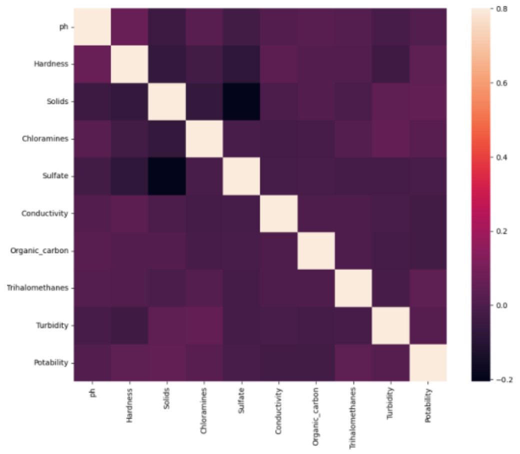


FIGURE 14. GRAPH OF CORRELATION BETWEEN FEATURES

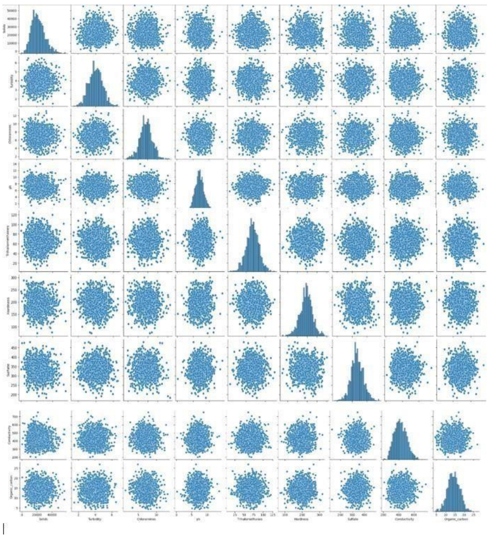
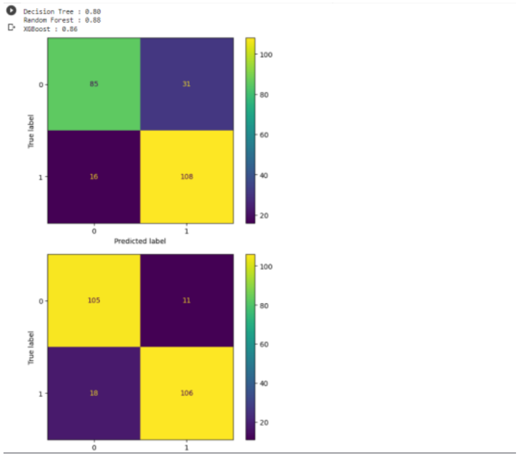


FIGURE 15. GRAPH OF PAIR PLOTS

algorithms were compared with Deep Learning using ANN, itcan be inferred and concluded that ML algorithms showed better performance as Deep Learning doesn't perform well on numerical, organised data place in rows and columns. DeepLearning algorithms perform well on unorganised and unstructured data and images.

IV. CONCLUSION

Intensive and comprehensive approach was performed to study the potability of water. In this study water qualityanalysis was carried out. During the first stage of the project,a software simulation was performed using 3 sensors tem- perature sensor, turbidity sensors and pH sensor on Proteus platform. Then in the second stage a hardware project was set up using these 3 sensors, Arduino and ESP8266 to test real-time various water samples and these values were stored in cloud server using Node MCU. This hardware can be implemented for real-time analysis of the various water samples and predict their potability. An extensive project wascarried out to explore further classification of potable water using various ML algorithms in the third and final stages of project.



```
Decision Tree : 0.80
Random Forest : 0.88
XGBoost : 0.86
```

FIGURE 16. CONFUSION MATRIX OF DECISION TREE AND RANDOMFOREST

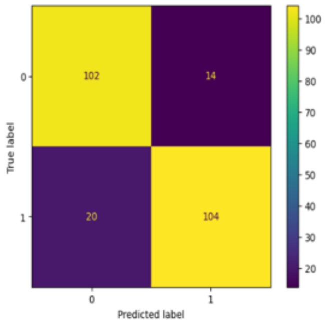
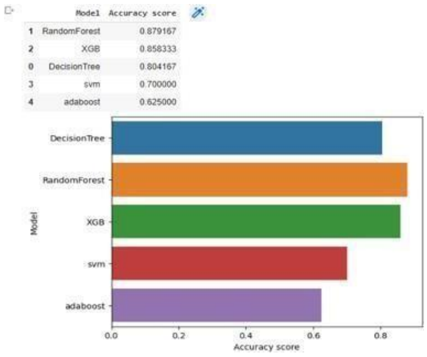


FIGURE 17. CONFUSION MATRIX OF XGBOOST



	Model	Accuracy score
1	RandomForest	0.879167
2	XGB	0.858333
0	DecisionTree	0.804167
3	svm	0.700000
4	adaboost	0.625000

FIGURE 18. ACCURACY SCORE OF ALL MACHINE LEARNING

Balancing the dataset using re-sampling and shuffling helped to increase the accuracy of the models and prevent bias towards the majority class. pH, temperature, and few other parameters were identified as important predictors of potable water classification based on their strong positive correlation with the target variable. The heat map prepared in data visualization shows no parameter is highly correlated to the other, from which it can be inferred that no parameter can be ignored while predicting the water potability. The Random Forest algorithm performed the best, achieving an accuracy of 88 percent (Fig 16,10), and had the highest precision,recall,F1 and MCC scores(Fig 10). The Decision Tree and XGBoost algorithms also performed well, achieving accuracies of 80 percent and 86 percent, respec- tively(Fig 16,17). The SVM and ANN algorithms performedpoorly, achieving accuracies of 70 percent and 68 percent respectively (Fig 19,21). KNN algorithms also showed poor performance 66 percent(Fig 22) but the Adaboost was the worst with 63 percent accuracy(Fig 20).

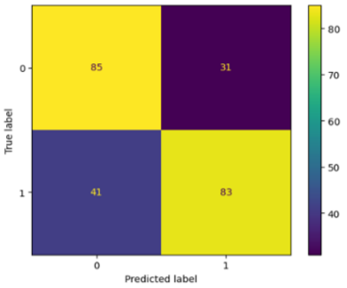


FIGURE 19. CONFUSION MATRIX OF SVM

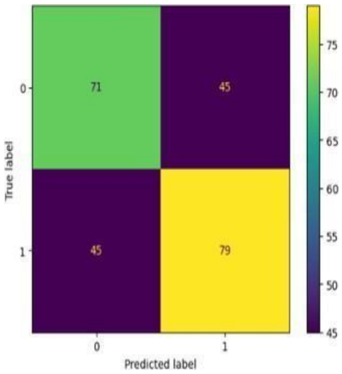


FIGURE 20. CONFUSION MATRIX OF ADABOOST

Performance of the ML techniques were also evaluated us-ing accuracy, precision, recall, F1 Score and MCC score(Fig 11), which reconfirmed the highest performance of RF algo- rithm. Performance of all the ML algorithms were compared against deep learning ANN, it was found all the tree based classifiers (ML algorithms) outperformed the Deep Learningalgorithm (ANN), with RF showing the best accuracy of 88 percent. It can also be reasonably concluded and deduced that deep learning algorithms have limited performance on tabular and numerical data which is not linearly separable.

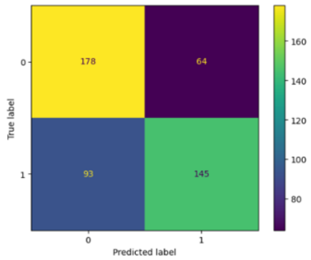


FIGURE 21. CONFUSION MATRIX OF ANN

DL agorithms are superior for images and text.These results suggest that the Random Forest algorithm is a promising approach for classifying potable water and can be used for future research in this area. As a future scope the IOT based hardware system can explored to testing real-time water samples and analysing their potability.

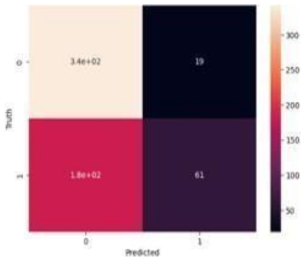


FIGURE 22. CONFUSION MATRIX OF KNN

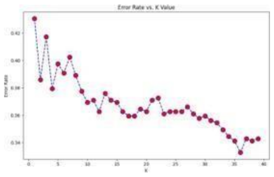


FIGURE 23. GRAPH OF ERROR RATE VS K.VALUE

REFERENCES:

1. *Nayla Hassan Omer* Water quality Published 16 October 2019 in Environmental Science, ’Water Quality Parameters’.
2. *Manisha Koranga, Pushpa Pant, Tarun Kumar, Durgesh Pant, Ashutosh Kumar Bhatt, R.P. Pant* Materials Today: Proceedings 57 (2022) 1706–1712 Contents lists, ’Efficient water quality prediction models based on machine learning algorithms for Nainital Lake,Uttarakhand’.
3. *Sandeep Bansala and G.Geethab* THEORETICAL PRINCIPLES OF WATER PURIFICATION AND TREATMENT TECHNOLOGY Journal of Water Chemistry and Technology, 2020, Vol. 42, No. 5, pp. 321 “A Machine Learning Approach towardsAutomatic Water Quality Monitoring”.
4. *Sudhakar Singha, Srinivas Pasupuleti, Soumya S.Singha, Rambabu Singh, Suresh Kumar* Chemosphere, Elsevier Published on 15 March 2021 “Prediction of groundwater quality using efficient machine learning technique”.
5. *LIANG KUANG, PEI SHI, CHI HUA, BEIJING CHENAND HUI ZHU*IEEE Access in SPECIAL SECTION ON TOWARDS SMART CITIES WITH IOT BASED ON CROWDSENSING, VOLUME 8, 2020 “An Enhanced Extreme Learning Machine for Dissolved Oxygen Predictionin Wireless Sensor Networks”.
6. *S.Angel Vergina, Dr.S.Kayalvizhi, Dr. R.M. Bhavadharini, Kalpana Devi.S* European Journal of Molecular Clinical Medicine, 2020, Volume 7, Issue 8, Pages 2035-2041“AReal Time Water Quality Monitoring Using Machine Learning Algorithm”.
7. *Hamza Khurshid, Rafia Mumtaz, Noor Alvi, Ayesha Haque, Sadaf Mumtaz, Faisal Shafait, Sheraz Ahmed,Muhammad Imran Malik Andreas Dengel* Springer on Published 28 January 2022,in EnvironmentalMonitoring and Assessment 194, Article number 133 (2022) “Bacterial prediction using internet of things (IoT) and machine learning”.
8. *Mourade Azrour, Jamal Mabrouki,Ghizlane Fattah, Aze- dine Guezzaz Faissal Aziz*Springer on Published 26 August 2021 in Modeling Earth Systems and Environment volume 8, pages 2793–2801 (2022) “Machine learning algorithms for efficient water quality prediction”.
9. *P.Kirankumar, G. Keertana, S U Abhinash Sivarao, B. Vijaykumar, Sh. Chetan Shah* IEEE Xplore International Conference on Intelligent Systems,Smart and Green Technologies on Published02 March 2022 “Smart Monitoring and Water Quality Management in Aquaculture using IOT and ML”.
10. *Haeng Yeol Oh, Myeong-Hun Jeong, Seung Bae Jeon, Tae Young Lee†, Gun Kim and Minkyo Youm* Journal of WaterProcess Engineering, ELsevier 48 (2022) 102920 Published on 31 May 2022 “Sea Water Quality Estimation Using Machine Learning Algorithms”.
11. *Nida Nasir, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, Ahmed Al- Shamma’a* Journal of Coastal Research on published on 2021 in Journal of Coastal Research “Water quality classification using machine

learning algorithms”.

12. *Mohamed Djerioui, Mohamed Bouamar, Mohamed Ladjal, Azzedine Zerguine* Springer Published online: 19 April 2018 in Arabian Journal for Science and Engineering (2019) 44:2033–2044 “Chlorine Soft Sensor Based on Extreme Learning Machine for Water Quality Monitoring”.
13. *S. Aishwarya and J. Abinaya* International Journal of Engineering Research Technology (IJERT) on Published on Special Issue - 2018 in Volume 6, Issue 03“Real Time Water Quality Monitoring Using WSN”.
14. *Hye Won Lee, Min Kim, Hee Won Son, BaehyunMin,Jung Hyun Choi* ELSEVIER Journal of Hydrology: Regional Studies 41 (2022) 101069 on Published on 7 April 2022 “Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea”.
15. *Md Galal Uddin, Stephen Nash, Mir Talas Mahammad Diganta, Azizur Rahman, Agnieszka I. Olbert* ELSEVIER Journal of Environmental Management Published on 19 August 2022 “Robust machine learning algorithms for predicting coastal water quality index”.
16. *Illa Iza Suhana Shamsuddin, Zalinda Othman and Nor Samsiah Sani* MDPI Published on 20 September 2022 in Water 2022, 14, 2939 “Water Quality Index Classification Based on MachineLearning: A Case from the Langat River Basin Model”.
17. *Yue Ma,Kaishan Song, Zhidan Wen, Ge Liu, Yingxin Shang, Lili Lyu,Jia Du, Qian Yang, Sijia Li, Hui Tao and Junbin Hou* IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 14, Published on 2021 “Remote Sensing of Turbidity for Lakes in Northeast China Using Sentinel-2 Images With Machine Learning Algorithms”.
18. *Water Quality* “https://www.kaggle.com/datasets/adityakadiwal/water-potability/code ”.
19. *Water quality Dataset* https://www.kaggle.com/datasets/adityakadiwal/water-potability

LAVANYA SINGH

She is pursuing a Bachelor of Technology degree in Electronics and Communication at the Vellore Institute of Technology in Chennai, Tamil Nadu. She completed an internship in VLSI at Maven Silicon. Deep learning, machine learning, big data and wearable technology are some of her research interests. She is a part of NSS VIT Chennai. She successfully finished a robotics and automationworkshop. She was a member of one of VIT Institute’s top 20 teams in Flipkart Grid 2.0.

SHAIK ABDUL KALAM

He is currently pursing Bachelor of Techologywitha specialisation major in Electronic and communication, and minor in computer science, from VIT Chennai, Tamil Nadu. He is a Student coordinator of IEEE RAS club from VIT Chennai.He is currently working as an Intern at Mindtree.

DR SIVAKUMAR S

Dr. S. Sivakumar,from Chennai,India. In 1999, he graduated with a B.E. in Electronics and Commu- nication Engineering from Bharathiyar University in Coimbatore. In 2006, he earned his Master of Engineering in Communication Systems from Anna University in Chennai. In 2015, he was awarded his Ph.D. from Anna University in Chen- nai. In the School of Electronics Engineering at Vellore Institute of Technology in Chennai, he is now an Associate Professor Grade 2. He is an ISTE Life Member. His research interests include wireless sensor networks, the internet of things, digital signal processing, image processing, and computer networking. He has around 21 years of teaching experience.