IoT based prediction of water quality index for farm irrigation

Agriculture sector of Indian economy in which more than half of the population is involved, contributes to less than quarter of the GDP. With advancement in ICT, tools and techniques can be developed that can help in analyzing and automating various phases of farming for improving productivity. This work focuses on analysing the quality of irrigation water and developing a model for prediction of Irrigation Water Quality Index(IWQI) based on Salinity and Sodicity. Development of IWQI can save time and cost of lab tests for irrigation water. Five parameters of water Na, Cl, EC, HCO and SAR are measured using which the IWQI is calculated. These five parameters are further reduced to three parameters using correlation analysis and a classification model for prediction of water quality class is developed using various classification techniques. Best result is obtained by Random Forest Classifier followed by Gradient Boosting and Neural Network Classifier. The classification model can be used in IoT based farming systems for preventing salinity based damage to the crops.

Irrigation Water Quality Index, Precision Agriculture, Water salinity, Sensor Data

# Introduction

In a survey conducted in 2018, it was found that 50% of the workforce in India is involved in the agriculture sector, however its contribution is only 16% to the GDP.There exists a need to increase the efficiency of each stage in farming and at a cheap cost so that it is affordable. With improvement in technology, these needs can be addressed using innovative solutions like the Internet of Things(IoT). This technology allows us to use sensors which in turn can be used to obtain large datasets for physical and chemical characteristics of soil, water and weather. This data can be analysed and machine learning tools can be developed that can help in proper utilization of resources and increase the overall efficiency of crop production.

We require proper water quality standards for irrigation to improve crop yield and help maintain the soil productivity. Deficiency or excess of nutrients in soil also depends on water used for irrigation which effects the plant growth. However, most existing methods for water quality standards are for drinking water quality which is not suitable for plants and some methods which do exist for irrigation water quality use high number of parameters which ist’t economically feasible for a farmer.

In this paper, we develop a model for prediction of Irrigation Water Quality Index(IWQI) class based on the degree of sodicity and salinity of water. The contribution of our work are as follows:

* We convert 5 parameters to quality measurement values which are used to calculate IWQI. We construct IWQI classes using IWQI values for irrigation water.
* We reduce 5 parameters to 3 parameters using correlation analyis to save costs of sensors required to implement the proposed work. These parameters are used to classify IWQI using seven classification techniques which are Support Vector Classifier, Neural Networks, Gradient Boosting , Random Forest, Decision Tree, Bagging and Naive Bayes classifier.
* We evaluate and choose the best performing classification algorithm to obtain IWQI class for water sample with 3 parameters.

The rest of the paper is organized as follows: Section [2](#section:realtedWork) describes other works related to this paper, section [3](#section:preliminaries) discusses the preliminary topics, Section [4](#section:materialAndMethods) describes the database used and algorithm of the model, Section [5](#section:expAndRes) describes the results and finally Section [6](#section:conclusionAndFutureWork) gives the conclusion and future work.

# Related Works

With advancement in IoT, many new tools and techniques have been developed for agriculture sector. The term Precision agriculture was developed in which various types of data such as time series, geolocation, sensor etc. are gathered and analysed to improve the efficiency of agricultural practices. Irrigation is one of the important phases of agriculture which involves adding water to soil after plantation which needs proper analysis of chemicals in water and quality of water for proper plant growth.

Rajesh et al proposed an automated system that measures the moisture, temperature and humidity of farms measured using sensors to determine the optimal water volume required for Irrigation. Mark et al. studied the effect of saline water on plants and proposed methods for using them for irrigation while Shrivastav et al. studied the effect of salinity on soil nutrients which decreases with increase in salinity due to its ability to reduce soil permeability.

Irrigation Water quality Indices(IWQIs) are one way to aggregate various parameters of water samples into a single value and determine their suitability for irrigation. Meireles et al developed a WQI for irrigation water (which mainly focused on salinity and sodicity of water) using factor and principal component analysis to reduce thirteen parameters of water samples to just five parameters. Sayiter et al used an Irrigation Water Quality Index(IWQI) using Sodium Adsorption Ratio (SAR), Kelly Index(KI), sodium percentage and Permeability Index(PI) and developed a model using Neural Networks for prediction of IWQI. Singh et al uses index based approach for predicting water quality based on 12 parameters to classify water into 5 categories: excellent, good, medium, bad, and very bad. For performing this classification, they use multiple criteria decision analysis (MCDA) approach, indicating the importance of IWQI for managing water quality index. Hossam et al used various dimensional reduction techniques such as Principal Component Analysis(PCA) and Factor Analysis(FA) to reduce 22 water quality parameters to 7 parameters:Sodium, Zinc, pH, Bicarbonate, Astantine, Boron and Nitrate and derived their weights using PCA. In the proposed work, we use the the IWQI developed by Meireles et al and develop a classification model to predict the IWQI class.

# Preliminaries

## Salinity Based Irrigation Water Quality

Calculation of water quality for irrigation is necessary to study its effect on plant and soil health. Water Quality Index (WQI) is used for representing the overall quality of water. It is the aggregate of various chemical and physical parameters of water and is calculated by multiplying relative weights of parameters with parameter value. The parameters chosen should be able to be measured in all the water sources used for irrigation. In this work, we use the Irrigation Water Quality Index (IWQI) based on salinity and sodicity of water. The following subsections describe the water quality parameters, their weights and formula for calculating IWQI.

### Parameters for IWQI

The IWQI used in this work was formulated by Meireles et al who reduced 13 parameters of water using Factor Analysis(FA) and Principal Component Analysis(PCA) to 5 parameters which are: Sodium, Chloride and Bicarbonate ion concentration, Sodium Absorption Ratio (SAR) and Electrical Conductivity(EC) . Electrical conductivity is generally measured by passing current through solutions and the value of SAR is given by the combination of Calcium, Magnesium and Sodium ion concentration whose formula is given by Equation [[equation:sar]](#equation:sar). It is the primary indicator of sodicity of water which affects the permeability of soil and water infiltration rate. Both cases of water infiltration rate being too high or too low affects the growth of plants. The weights of the 5 parameters, which represents their contribution to water quality are converted to relative weights and are mentioned in Table [1](#table:wqiParams).

Here, : Concentration of Sodium ion : Concentration of Calcium ion : Concentration of Magnesium ion

Weights for IWQI parameters

|  |  |
| --- | --- |
| **Parameters** | **Weights** |
| Electrical Conductivity(EC) | 0.211 |
| Sodium(Na) | 0.204 |
| Bicarbonate(HCO) | 0.202 |
| Chloride(Cl) | 0.194 |
| Sodium Absorption Ratio (SAR) | 0.189 |

[table:wqiParams]

### Quality Measurement Values and IWQI

The value of parameters measured for water sample should neither be too high nor too low for optimal growth of plants due to which the parameters are normalized(between 0 to 100) to quality measurement values () according to limits shown in Table [3](#table:qvalues) and Equation [[equation:calcQ]](#equation:calcQ). The values obtained are multiplied with relative weights of their respective parameters and finally added to obtain IWQI according to Equation [[equation:calcWQI]](#equation:calcWQI) whose value is between 0 and 100. IWQI range is divided into various classes and each class represents the salinity characteristic.

Here, : quality measurement values, : maximum value in particular quality measurement class, : amplitude of quality measurement class, : value of parameter x, : minimum value of parameter x in quality measurement class, : It is the class amplitude to which the parameter belong

Here, : quality measurement value , : weight of parameter from Table [1](#table:wqiParams)

Classification of water sample based on IWQI

|  |  |  |
| --- | --- | --- |
| **IWQI** | **Soil** | **Plant** |
| 85–100 | Can be used for any kind of soil | Most plants won’t be affected |
| 70–85 | Can be used on soil with moderate permeability | Avoid use in plants with very low salt tolerance |
| 55–70 | Can be used on soils with moderate to high permeability | Avoid in plants with low salt tolerance |
| 40–55 | Can be used on soils with high permeability without dense layers | Used mainly in plants with high salt tolerance. Plants with moderate salt tolerance can be used with some control practices |
| 0–40 | Use for irrigation should be avoided | Avoid for all plants |

[table:classifyWQI]

Limiting values for quality measurement

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **EC** | **SAR** |  |  |  |
| 85-100 | 0.2 | 2 | 2 | 1 | 1 |
| 60-85 | 0.75 | 3 | 3 | 4 | 1.5 |
| 35-60 | 1.50 | 6 | 6 | 7 | 4.5 |
| 0-35 | or EC | or SAR | or | or | or |

[table:qvalues]

## Classification Algorithms

### ANN

In ANN, we have an input layer, hidden layers and output layer each consisting of neurons in it. Neurons receive some input and give an output after applying an activation function. The error is backpropagated to network for adjusting weights. This classifier is also known as MLP(Multilayer Perceptron Classifier).

### SVM

This classification algorithm is very useful as it tries to maximize the margin between any two classes which makes the decision hyperplane more accurate.

### Gradient Boost Classifier

It is a classifier which uses loss function, weak learners and an additive component to reduce the loss. This algorithm uses decision trees as its weak learners. It is quite powerful but is prone to overfitting.

### Random Forest Classifier

It is an ensemble learning method which uses multiple trees. Here each classifier gives its own output and the output which is chosen by more number of classifiers becomes the final output.

### Decision Tree

It is a classifier which uses a tree-like structure where at each node a decision is made according to the feature in the dataset and final output is given by the leaf node.

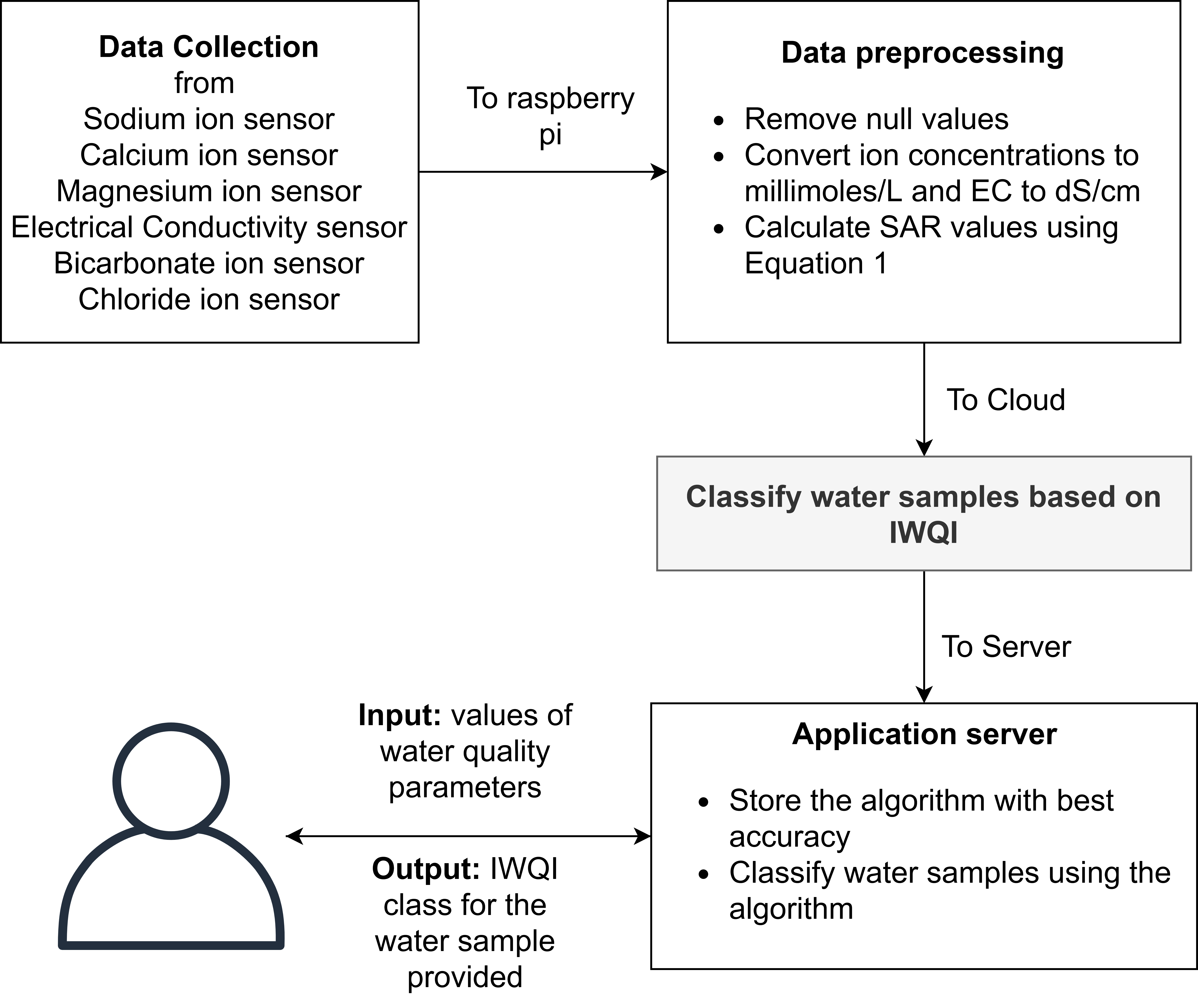
### Naive Bayes

It is a simple classifier which uses Bayes theorem with an exception that features do not depend on each other. This allows us to easily implement and use the algorithm for classification but its results are not reliable.

### Bagging Classifier

This algorithm uses weak learners like decision trees on a random portion of its dataset, using the results from those trees to produce a final result.

# Proposed Work



Overall architecture

In this section, we propose a classification model for prediction based on the Irrigation Water Quality(IWQI) class described in section [3.1.2](#subsubsection:calcIWQI). The model can be incorporated into agricultural IoT systems for analysing water samples for irrigation purposes as shown in Fig. [1](#fig:overallArchitecture). The sensor network will be used to measure the values of various water quality parameters namely electrical conductivity and ion concentrations for Sodium, Magnesium, Calcium, Bicarbonate and Chloride. Data from sensor network is sent to a local machine like Raspberry pi that involves processing the data and sending it to the cloud. The cloud analyses the data by predicting the relationship between water quality parameters and the IWQI classes. Result of classification can be saved to an application server than can interact with users through web or mobile applications. In this work, we have focused on the data analysis part which involves developing the classification model described in section [4.2](#subsection:iwqiClassification) used the water quality data from major ions dataset that is described in section [4.1](#subsection:datasets).

## Dataset - Major ions dataset

The dataset used in our work was obtained from US Geological Survey of Brackish Groundwater. It consists of 66 thousand rows and contains values for concentration of dissolve solids, metals, nutrients, ions and physical properties such as conductivity out of which only EC, Cl, Na, HCO, Ca and Mg are used for IWQI. EC stands for electrical conductivity which is measured in milliSiemens/cm. Cl, Na, HCO, Ca and Mg represents Chloride, Sodium, Bicarbonate, Calcium and Magnesium ion concentration respectively. These concentrations were measured in mg/L which we converted to millimoles/L. EC needs to be in dS/cm so it was divided by 100. We use Ca, Mg and Na to calculate SAR using Equation [[equation:sar]](#equation:sar).

## Irrigation Water Quality Index based classification

In this work, we develop a classification model and apply it on major ions dataset. This dataset consists of all the necessary features required to obtain the IWQI described in Section [3.1.1](#subsubsection:WaterQualityParameters) and the model is built using these three steps: converting given values to quality measurement values, calculating irrigation water quality index(IWQI) using calculated values and relative weights of each parameter, assigning to a class according to ranges given in Table [2](#table:classifyWQI) and finally using different classification techniques and choosing the best of them.

### Quality Measurement Values

The dataset consists of parameters with different units. Also, the value of any parameter being abnormally high or low decrease the quality of water. Hence, we need to normalize the values according to predefined limits of parameters. This is done by converting values of parameters to quality measurement values () that can be obtained using Equation 2 and Table 3. The quality measurement value () is a dimensionless quantity which can then be combined with other parameters without considering units. The Table [4](#table:numQualityMesurementInrange) indicates the number of values present in specific ranges of for the dataset used. Here, we can see that majority samples have values of EC in the range 85-100, bicarbonate in the range 35-85 and Cl, Na and SAR in the range 0-35. This indicates that most of the samples have optimal value of EC and either excess or deficit value of Cl, Na and SAR.

Number of Quality measurement values in each range

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Q\_HCO** | **Q\_EC** | **Q\_Cl** | **Q\_Na** | **Q\_SAR** |
| 0-35 | 5540 | 6580 | 32700 | 31400 | 32600 |
| 35-60 | 9950 | 0 | 0 | 0 | 794 |
| 60-85 | 20000 | 3500 | 0 | 2780 | 1950 |
| 85-100 | 2470 | 27900 | 5280 | 3810 | 2680 |

[table:numQualityMesurementInrange]

### Calculating IWQI and developing a classification model

Using the quality measurement values calculated in section [3.1.2](#subsubsection:calcIWQI), we multiply them by their relative weights mentioned in Table [1](#table:wqiParams) and add them to obtain IWQI according to Equation [[equation:calcWQI]](#equation:calcWQI). The samples are further classified into quality classes according to quality ranges provided in Table [2](#table:classifyWQI). Correlation matrix of quality parameters which was obtained using Pearson’s Correlation is given in Table [5](#table:correlationMatrixOfIons&IWQI).The following observations are made:

1. Cl is highly correlated with EC and Na.
2. Na is highly correlated with all the parameters.
3. EC is highly correlated with all the parameters except SAR.
4. HCO is highly correlated with EC.
5. IWQI is highly correlated with Cl, Na and EC.

$$r = \frac{\sum(U\_i - \Bar{U})(V\_i - \Bar{V})}{\sqrt{\sum(U\_i - \Bar{U})^2(V\_i - \Bar{V})^2}}
\label{equation:correlationMatrixOfIons&IWQI}$$

Here, r: correlation coefficient U: value of U-variable in the sample $\Bar{U}$: mean value of U-variable V: value of V-variable in the sample $\Bar{V}$: mean value of V-variable

correlation matrix for various parameters for water samples

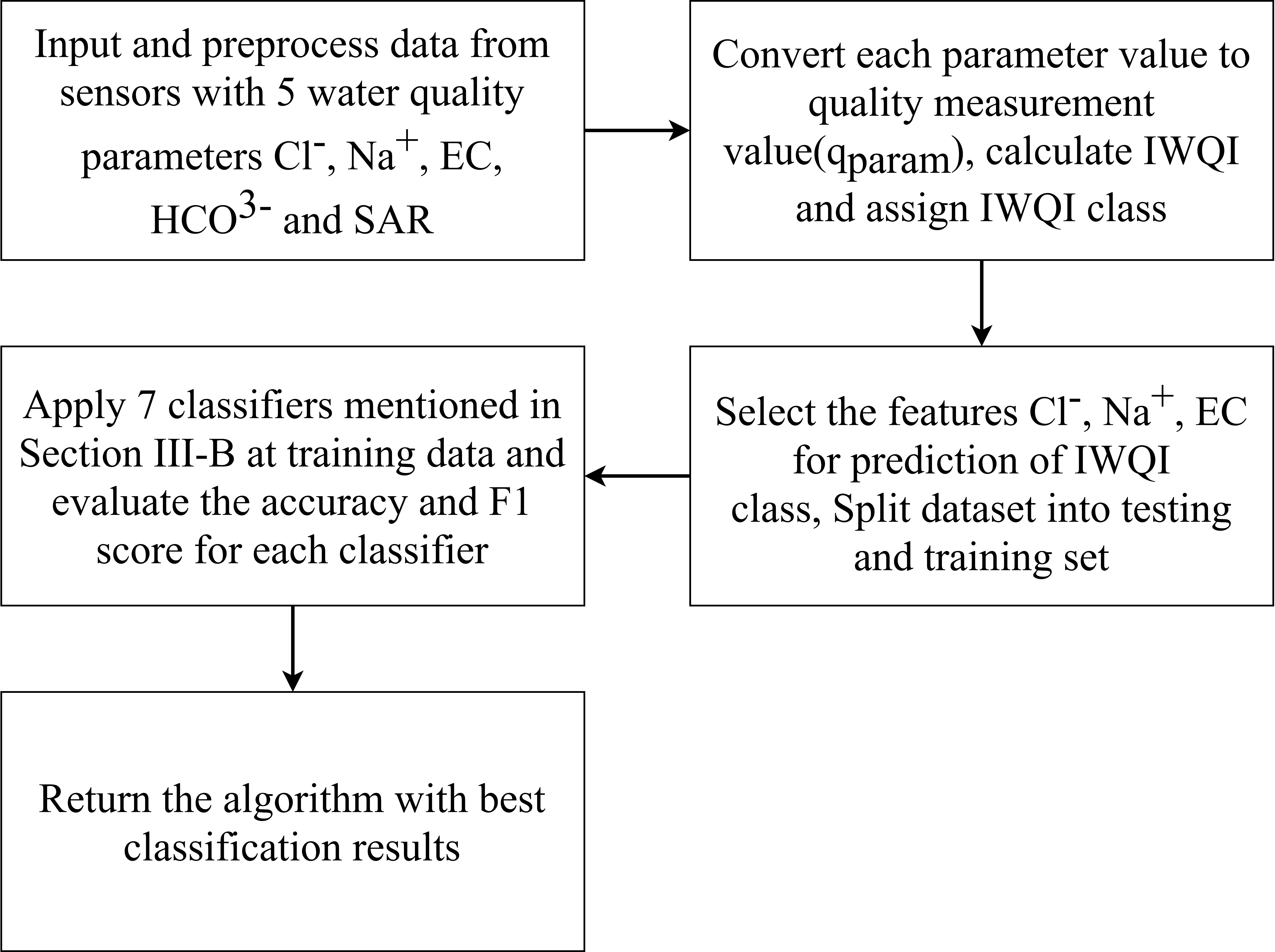
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Cl | Na | EC | HCO | SAR | IWQI |
| Cl | 1 | 0.467 | 0.506 | 0.159 | 0.083 | 0.44 |
| Na | 0.467 | 1 | 0.482 | 0.297 | 0.56 | 0.607 |
| EC | 0.506 | 0.482 | 1 | 0.835 | 0.018 | 0.423 |
| HCO | 0.159 | 0.297 | 0.835 | 1 | 0.018 | 0.293 |
| SAR | 0.083 | 0.56 | 0.018 | 0.018 | 1 | 0.302 |
| IWQI | 0.44 | 0.607 | 0.423 | 0.293 | 0.302 | 1 |

[table:correlationMatrixOfIons&IWQI]

We used different combination of parameters according to their correlation with IWQI value as features to the classification model in which Na, Cl and EC were found to be most suitable for prediction IWQI class. The classification result obtained using the three parameters are almost similar to that obtained using all of the five parameters. Therefore, correlation analysis is sufficient and other feature reduction techniques are not required. Choosing three parameters saves us the cost of measuring HCO and SAR. The classification algorithms mentioned in Section [3.2](#subsection:classificationAlgos) are used to predict quality classes and the method having best accuracy and F2 score is selected for further use. Multiple classification techniques have been used to cover different traditional, ensemble and deep learning based techniques whose performance can help in predicting the most suitable type of classification technique for the given dataset. The overall algorithm is mentioned in algorithm [[algorithm:classifyIWQI]](#algorithm:classifyIWQI) and flowchart in Fig. [2](#fig:classificationFlowchart).

[algorithm:classifyIWQI] Water quality dataset with 5 parameters Cl, Na, EC, HCO and SAR

algorithm with best score Calculate quality measurement values(q) for each parameter. Calculate IWQI using the Equation [[equation:calcWQI]](#equation:calcWQI) and assign classes from Table [2](#table:classifyWQI) Choose features of dataset as: features Cl,Na,EC Choose labels of dataset as:IWQI Perform train test split on the dataset Apply the 7 classifiers mentioned in Section on the training set and evaluate on the testing set using accuracy and F1 score. algorithm with best score



Flowchart for classification of water sample based on IWQI

# Experiments and Results

This section contains the simulation settings for implementing the classification model in [5.1](#subsection:toolsAndSimulationSettings), the evaluation methods for the model [5.3](#subsection:evaluationMetrics) and the results of algorithms in section [5.4](#X17877620fcd5bc15e1f419e50e99496519bd4b9).

## Tools and Simulation Settings

The IoT system for classifying water samples based on IWQI values that is shown in Fig. [1](#fig:overallArchitecture) requires measurement of various water quality parameters mentioned in section [3.1.1](#subsubsection:WaterQualityParameters). As we are focused on the classification part, we obtain the parameters from the Major ions dataset described in section [4.1](#subsection:datasets). The algorithm was implemented in Python 3 where data preprocessing is performed using Pandas and Numpy . Various classification algorithms are applied using Scikit-learn. The parameter settings for these algorithms are mentioned in section [5.2](#subsection:parameterSettings).

## Parameter Settings

The parameter settings for all the classification models are given in Table [6](#table:parametersClassifiers) and the parameters are described below:

### Decision Tree

criterion: function to evaluate split quality, splitter: method used to split at each node, min\_sample\_split: least amount of samples required to split a non-leaf node

### Naive Bayes

alpha: the additive smoothing parameter, class\_prior: estimated probability for new class prediction, class\_count: number of samples under each class

### Gradient Boosting

loss: function to optimize, learning\_rate: for changing contribution of trees to the result, N\_estimators: levels of boosting

### Random Forest

n\_estimators: Amount of tree before voting, criterion: function to evaluate split quality, base\_estimator: base classifier to create sub estimator

### Support vector classifier

C: Regularization, decision\_shape: decision function’s shape, tol: stopping tolerance

### Bagging Classifier

base\_estimator: base classifier used to fit on random subdomains, n\_estimator: count of the base\_estimator

### MLP Classifier

hidden\_layer\_sizes: array representing sizes of hidden layers, activation: output function of hidden layer, learning\_rate: step size of gradient descent

Parameter Settings for Classifiers

|  |  |
| --- | --- |
| **Classifier** | **Parameter Settings** |
| Bagging Classifier | base\_estimator:SVC, n\_estimators:10 |
| DecisionTree | criterion: gini, splitter: best, min\_samples\_split: 2 |
| Naive Bayes | alpha: 1.0, class\_prior: None, class\_count: [ 7059. 15826. 5057. 2055. 402.] |
| Gradient Boosting | loss: deviance, learning rate: 0.1, n\_estimators: 100 |
| Random Forest | n\_estimators: 100, criterion: gini, base\_estimator\_: DecisionTreeClassifier |
| SVM | C: 1.0, decision\_function\_shape: ’ovr’, tol:0.001 |
| MLP | hidden\_layer\_sizes:128, activation=’relu’, learning\_rate=0.001 |

[table:parametersClassifiers]

## Evaluation Metrics For Classification Model

For testing our classifier, we use precision, accuracy, recall and F1 score. These metrics are calculated to test the reliability and robustness of classifiers used . These metrics are defined in Section [5.3.1](#subsubsection:accuracy), [5.3.2](#subsubsection:precision), [5.3.3](#subsubsection:recall) and [5.3.4](#subsubsection:f1score). The equations for the same are given at [[equation:accuracy]](#equation:accuracy), [[equation:precision]](#equation:precision), [[equation:recall]](#equation:recall) and [[equation:f1Score]](#equation:f1Score).

### Accuracy

It tell us how much data can be predicted accurately from a given classifier. This gives us an idea about the bulk of data which is correctly classified.

### Precision

It tells us how many predictions are correct out of all observations saying output belongs to a particular class.

### Recall

It tells us how many correct predictions were made for a class out of all the actual observations for a given class.

### F1 Score

It can be defined as harmonic mean of precision and recall, can be used independently for evaluation of classifiers.

Here, TPos: Number of observations which are correctly classified for positive class in binary classification TNeg: Number of observations which are correctly classified for Negative class in binary classification FPos: Number of observations which are incorrectly classified for positive class in binary classification FNeg: Number of observations which are incorrectly classified for negative class in binary classification

## Results of Classification based on IWQI

Algorithm [[algorithm:classifyIWQI]](#algorithm:classifyIWQI) is applied to major ions dataset with Electrical Conductivity, Cl and Na as features and IWQI class as labels and results are evaluated using the metrics mentioned in Section [5.3](#subsection:evaluationMetrics), details of which are mentioned in Section [4.2.2](#subsubsection:classificationBasedOnIWQI). For the major ions dataset, the number of water samples in IWQI range 85-100 is around 1.3% of the total samples which affected the results of correctly classifying samples belonging to that range. This however does not adversely affects the overall result. Random Forest performed the best with an accuracy of 86.9% followed by Gradient Boosting and Neural Networks with accuracy of 85.8% and 84.6% respectively. Naive Bayes classifier performed the worst with accuracy of 52.6%. The results of all the algorithms are given in Table [7](#table:classificationResults). We also performed five-fold cross validation and computed the accuracies whose results are in Table [8](#table:fiveFoldCrossValidation). This concludes that all of the methods except Naive Bayes perform well in classifying the water samples.

Classification results for algorithm [[algorithm:classifyIWQI]](#algorithm:classifyIWQI)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methods** | **Accuracy** | **Precision** | **Recall** | **F1** |
| DecisionTree | 0.832 | 0.728 | 0.740 | 0.733 |
| Naive Bayes | 0.526 | 0.105 | 0.200 | 0.138 |
| Gradient Boosting | 0.858 | 0.762 | 0.757 | 0.760 |
| Random Forest | **0.869** | **0.790** | **0.765** | **0.776** |
| SVM | 0.845 | 0.762 | 0.711 | 0.726 |
| Bagging | 0.813 | 0.750 | 0.664 | 0.690 |
| MLP | 0.846 | 0.734 | 0.743 | 0.738 |

[table:classificationResults]

Five fold cross validation scores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Fold 1** | **Fold 2** | **Fold 3** | **Fold 4** | **Fold 5** |
| Bagging Classifier | 0.829 | 0.823 | 0.814 | 0.830 | 0.813 |
| Decision Tree | 0.833 | 0.836 | 0.828 | 0.823 | 0.839 |
| Naive Bayes | 0.518 | 0.518 | 0.518 | 0.518 | 0.519 |
| Gradient Boosting | 0.862 | 0.866 | 0.869 | 0.859 | 0.862 |
| Random Forest | **0.866** | **0.870** | **0.870** | **0.862** | **0.871** |
| SVM | 0.853 | 0.849 | 0.846 | 0.841 | 0.848 |
| MLP | 0.847 | 0.853 | 0.857 | 0.847 | 0.849 |

[table:fiveFoldCrossValidation]

# Conclusion and Future Work

Analysis of irrigation water quality can help in increasing agricultural productivity and prevent damage to plants. In this work, we developed a classification model for prediction of Irrigation Water Quality Index (IWQI) class. The IWQI used here is an aggregate of five parameters that are SAR, EC, Chloride, Bicarbonate and Sodium ion concentration respectively. For classification, we used seven classification algorithms and selected three out of five parameters: EC, Cl and Na for predicting IWQI class. Random Forest Classifier performed the best followed by Gradient Boosting and Neural Networks. The water quality index used here is mainly representation of salinity of water and the dataset used here is of groundwater. Other parameters such as acidity, oxygen demand etc. can be used in addition to the parameters used here to develop an index covering more factors. Also, dataset covering different types of water bodies can be used to improve the model. The proposed work used here can be incorporated into IoT based agricultural systems for analysing irrigation water quality that will be faster and economically feasible compared to manual lab tests. It can also help in deciding the water sample according to crop and soil properties.