Noname manuscript No.
(will be inserted by the editor)

Predicting irrigation WQI and suggesting the best plants for a given water or soil sample

Dr. Rajesh Kumar Yadav · Adarsh Jha · Aditya Choudhary

Received: date / Accepted: date

Abstract Indian economy is highly affected by its agriculture sector in which more than half of the people are involved but its contribution to GDP is less than a quarter. One of the major reasons for the low productivity is inefficient agricultural practices. This paper is focused on predicting quality of irrigation water and suggesting plants based on soil and water properties. We used five properties of water quality: Electrical Conductivity(EC), Sodium Absorption Ratio(SAR), Chloride, Sodium and Bicarbonate ion concentration to calculate and further classify water samples based on irrigation water quality index(IWQI). Seven classification methods were used for the classification on major ions water dataset with Random Forest Classifier giving the best accuracy(86.9%) followed by Gradient Boosting and MLP network classifier. We also developed a model for suggesting plants based on Salinity and pH properties of soil and water. We applied the models on major ions dataset(for water quality index and plant suggestion based on IWQI) and LUCAS topsoil dataset (for suggesting plants based on soil properties).

Keywords Irrigation Water Quality · Precision Agriculture · Soil and Water salinity · Sensor Data

Dr. Rajesh Kumar Yadav
Assistant Professor
Department of Computer Science and Engineering, Delhi Technological University, New Delhi-110042, India
E-mail: rkyadav@dtu.ac.in

Adarsh Jha
Department of Computer Science and Engineering, Delhi Technological University, New Delhi-110042, India
E-mail: adarshjha_2k17co22@dtu.ac.in

Aditya Choudhary
Department of Computer Science and Engineering, Delhi Technological University, New Delhi-110042, India
E-mail: adityachoudhary_2k17co24@dtu.ac.in

1 Introduction

According to the economic survey of 2018 in India, around 50% of the workforce is involved in the agriculture sector[1]. The contribution of the sector to the GDP is only 16% which has reduced significantly from 50% in 1950. The decline is not limited to India and is observed in the rest of the world. This low productivity depends on many factors, one of the major one is wastage of farming resources, money and time. Also, the majority of people involved in farming are from rural areas who are poor and have insufficient knowledge regarding farming practices. Each phase of farming, be it preparation of soil, nutrient management, irrigation or harvesting, requires proper analysis of soil, water, weather, sunlight etc. for improving the productivity of crops. Study of these parameters requires lab tests which are expensive. Due to farmers being relatively poor, there is a need to develop methods which would find parameters cheaper than lab tests. With improvement in technology, various farming equipment have been developed aiming to increase productivity and proper utilization of resources. The use of the Internet of Things, which consists of various devices connected through a network interacting with computers to transfer data, is becoming feasible with the internet and computers getting faster and cheaper. Various sensors can be placed in a farming environment to obtain large, real-time datasets such as irrigation water(pH, Electrical Conductivity, concentration of various ions), soil(moisture, pH, nutrients)) , weather(temperature, humidity etc). These datasets can then be analyzed and used to train machine learning models that will help in proper utilization of resources, determining which crop could be suitable for a specific environment and increasing the overall efficiency of crop production.

In this work we contributed by predicting IWQI and suggesting plants based on a given water or soil sample. We do this by calculating water quality parameters, SAR(Sodium Absorption Ratio), Quality Measurement values and use them to predict IWQI and classify water samples according to it. We use several classification algorithms like Decision Tree, SVM, ANN, Gradient Boost Classifier, Random Forest Classifier, Decision Tree, Naive Bayes and Bagging Classifier to do so. We suggest plants for any water or soil value based on the pH and electrical conductivity ranges.

The rest of the paper is structured as follows: Section 2 focuses on works related to this paper. Section 3 discusses the preliminary works which we have used in our paper. Section 4 describes the models which we have constructed for water quality prediction, plant suggestion and various datasets on which we have applied our models. Section 5 describes the parameter settings for our model, evaluation metrics and results of applying our models on the datasets. Section 6 gives a conclusion to our work and describes the improvements and new methods that we are planning to apply in the future.

2 Related Work

With recent developments in IoT and cloud computing, many new and cheap sensors have come out that can help in agriculture. Also, many methods have been developed which make use of these equipment for efficient farming. Precision agriculture[2] is the process developed which involves gathering various time series, geolocation and individual data, processing and analyzing them, and making use of them for improving productivity in agriculture. Any step of agriculture where data can be gathered can come under precision agriculture. For example, in irrigation systems data can be gathered and analyzed to find the quality of water, determining and classifying soil and fertilizing it accordingly which will help in better fertilizing.

Analyzing irrigation water can help in preventing damage to crops. Orozco et al.[3] studied various samples of water for irrigation and developed a water quality index(WQI) measure for drinking water based on WHO guidelines. Lerios et al.[4] developed a method for prediction and classification of water according to parameters such as pH, EC etc for drinking water using various classifying techniques like Decision Tree[5], Random Forest[6], Gradient Boost[7], MLP[8], Bagging[9] and Naive Bayes[10]. These indices were not suitable for water used in irrigation purposes. Meireles et al.[3] uses Factor Analysis(FA) and principal component analysis(PCA) on 13 parameters of water such as Electrical Conductivity(EC), pH and various ion concentration to identify parameters that would majorly interfere in affecting salinity and toxicity of the soil and plant. Five parameters Electrical Conductivity, Sodium, Bicarbonate, Chloride and Sodium Absorption Ratio(SAR) were found to be having greater weights in FA and PCA. An Irrigation Water Quality Index(IWQI) was developed accordingly with values of the parameters normalized using q-value normalization. Irrigation involves watering the soil after sowing the seed for proper growth of crops. Vij. et al[11] proposes the use of wireless sensor networks for measuring various parameters such as temperature, moisture, water level, weather etc. , passes it through Support vector regression(SVR), Random Forest Regression algorithms to classify soil type and predict the amount of water required for irrigation. Soil parameters such as pH, Cation Exchange Capacity(CEC), Salinity, Macro and Micro nutrients etc. affects the growth of plants. Gupta et al.[12] explains how salinity of soil decreases the water holding capacity of soil and accelerates the water loss from the plants due to osmotic pressure.Also, due to salinity, there is accumulation of various ions in plant tissue that may cause physiological damage to plants and may even cause death of plants. This is known as ion toxicity. Shrivastav and Kumar[13] explains the effect of salinity on soil nutrients. Gentili et al.[14] studies the effect of soil pH on soil nutrients and plant growth by analyzing flower size, pollen production, biomass etc. for ragweed. Mark and Khalid[15] studies the effect of saline water and salt on plants and proposes methods for using them in crop production and listing various crops with their threshold values of salinity. In this paper, we classify water samples based on irrigation water quality index with the help of machine learning using fewer numbers of parameters than sug-

gested by Meirels et al. Although similar work has been done previously, but for drinking water quality which has no relation with irrigation water quality. We also analyze various water and soil samples and suggest plants based on their optimal conditions that are mainly based on Salinity and pH.

3 Preliminaries

3.1 Development of Irrigation Water Quality

Testing of water for irrigation purposes is required to prevent any damage to crops due to chemicals present in water. Different kinds of indices for water quality can be developed according to the type of damage one wants to prevent after irrigation of plants. Meireles et al. developed an irrigation water quality index for reflecting soil salinity, sodicity and toxicity. 13 parameters were measured for various water samples out of which 5 parameters contributing the most to sodicity and salinity were selected using Factor Analysis(FA) and Principal Component Analysis(PCA). An Irrigation Water Quality Index(IWQI) was developed to represent salinity and toxicity risk of water samples using a single value. Details of the index are mentioned in the Sections 3.1.1, 3.1.2 and 3.1.3.

3.1.1 Water Quality Parameters

As mentioned in Section 3.1, 13 parameters Electrical Conductivity, pH, Calcium ion concentration, Sodium ion concentration, Sodium Absorption Ratio(SAR) etc. were measured. Factor Analysis(FA) and Principal Component Analysis(PCA) were applied to the data in which Electrical Conductivity, SAR, Carbonate, Sodium,Chloride and Chloride ion concentration showed highest correlation(> 0.9) among them. Also, they carried the majority of factorial load in Factorial Analysis(FA) hence they were selected for developing water quality index. The weights for each parameter are based on variance of the first factor of the factorial load matrix[16]. All the weights were added and weight of each parameter was divided by the total weight to obtain the relative weights which are shown in Table 1.

Parameters	**Weights**
Electrical Conductivity(EC)	0.211
Sodium(Na^+)	0.204
Bicarbonate(HCO^{3-})	0.202
Chloride(Cl^-)	0.194
Sodium Absorption Ratio (SAR)	0.189

Table 1: Weights for IWQI parameters

3.1.2 Sodium Absorption Ratio(SAR)

SAR is a numerical value given by a combination of Sodium, Magnesium and Calcium ion concentration which is given by Equation 1. SAR indicates the amount of sodium related to calcium and magnesium. It is the primary indicator of sodicity of water whose higher values affect the permeability of soil and can lead to decreased water infiltration rate[17]. Water infiltration rate is defined by the rate with which water can enter the soil. Infiltration rate should neither be too high nor too low. Both cases affect the growth of plants.

$$SAR = \frac{Na^+}{\sqrt{\frac{Ca^{2+}+Mg^{2+}}{2}}} \tag{1}$$

Here,
Na: Sodium ion concentration
Ca: Calcium ion concentration
Mg: Magnesium ion concentration

3.1.3 Quality Measurement Values and IWQI

The weights obtained in Table 1 shows the contribution of each parameter towards the water quality index. They when multiplied by various parameters and summing will give IWQI. But, the value of a parameter being too high or too low will reduce the quality of water due to which parameters were normalized to quality measurement values(q_i) according to limiting values of each parameter which are shown in Table 3[16] using the Equation 2. The IWQI is calculated by multiplying relative weights with quality measurement values given in Equation 3 . Using IWQI, we can divide water samples into classes which are given in Table 2[16]. Each class describes the salinity characteristics of soil and plants that can handle water sample belonging to it.

$$q_i = q_{imax} - \frac{(x_{ij} - x_{inf}) * q_{iamp}}{x_{amp}} \tag{2}$$

Here,
q_i: quality measurement values,
q_{imax}: maximum value in particular quality measurement class,
q_{iamp}: amplitude of quality measurement class,
x_{ij}: value of parameter x,
x_{inf}: minimum value of parameter x in quality measurement class,
x_{amp}: It is the class amplitude to which the parameter belong

$$IWQI = \sum_{i=1}^{n} q_i w_i \tag{3}$$

Here,
$IWQI$: Irrigation Water Quality Index,

q_i: quality measurement value ,
w_i: weight of parameter from Table 1

IWQI	Restrictions	Soil	Plant
85–100	No restrictions (NR)	It can be used for most soils with low probability of solidification and salinization	Most plants won't be affected
70–85	Low restriction (LR)	Use for soil with fine texture or moderate permeability	Avoid use in plants with salt sensitivity
55–70	Moderate restriction (MR)	Can be used in soils with high or moderate permeability	Plants with moderate salt tolerance will be unaffected
40–55	High restriction (HR)	Can be used on soils with high permeability without layers of compaction.	It should be used to irrigate plants with moderate to high salt tolerance with special salinity control practices
0–40	Severe restriction (SR)	Use for irrigation under normal conditions should be avoided.	Avoided for all plants

Table 2: Classification of water sample based on IWQI

q_i	**EC**	**SAR**	Na^+	Cl^-	HCO^{3-}
85-100	0.2≤ $EC < 0.75$	2≤ $SAR < 3$	2≤ $Na^+ <$ 3	1≤ $Cl^- <$ 4	1≤ $HCO^{3-} < 1.5$
60-85	0.75≤ $EC < 1.5$	3≤ $SAR < 6$	3≤ $Na^+ <$ 6	4≤ $Cl^- <$ 7	1.5≤ $HCO^{3-} < 4.5$
35-60	1.50≤ $EC < 3$	6≤ $SAR < 12$	6≤ $Na^+ <$ 9	7≤ $Cl^- <$ 10	4.5≤ $HCO^{3-} < 8.5$
0-35	$EC < 0.20$ or EC ≥ 3.00	$SAR < 2$ or SAR≥ 21	$Na^+ < 2$ or $Na^+ \geq$ 9	$Cl^- < 1$ or $Cl^- \geq 10$	$HCO^{3-} < 1$ or $HCO^{3-} \geq 8.5$

Table 3: Limiting values for quality measurement

3.2 Optimal values for Electrical Conductivity and pH in plants

For correctly suggesting plants according to the water and soil conditions provided in [18] and [19] we needed criteria for prediction. This was provided by [20] and [21]. We constructed our own Table 4 based on the ranges provided in [20] and [21] and these ranges are used to correctly classify plants according to the given water and soil sample.

Plant	Lower limit of pH	Upper Limit of pH	EC_tolerance
Garden Beet	6.5	8	8
Potatoes	5	5.5	4
Corn	5.5	6	4
Barley	5.5	6	8
Wheat	5.5	6	8
Carrots	5.5	6	2
Onions	6	7	2
Strawberry	5.5	6.5	2
Peas	6	7.5	2
Beans	6	7	2
Cabbage	6	7.5	6
Tomato	5.5	7.5	6
Broccoli	6	7	6
Asparagus	6	8	8
Spinach	6	7.5	8
Sunflower	6.5	7.5	8
Kochia	6	6.5	16
Sugar beet	6	7	16
Safflower	4	8	8
Fall rye	4.5	8	8
Oats	4.5	8	6
Yellow Mustard	5	8	6
Flax	5	7	6
Canola	5.5	6.5	6

Table 4: Optimal range of pH and Electrical conductivity values for plants

3.3 Classification Algorithms

3.3.1 ANN

Artificial Neural Networks are inspired by the structure of our brain where dendrites receive the message which is passed through axon[8]. In ANN, neurons are responsible for receiving the input and producing the output after applying the activation function. Each neuron has a weight which increases or decreases as the learning proceeds. Typically, neurons are aggregated into layers. These layers transform the given input, finally producing and output which gets tested with the actual output and the error gets back propagated.

3.3.2 SVM

SVM presents one of the most robust prediction methods[22]. Its objective is to find a hyperplane with maximum margin separation which distinctly classifies data points. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. It can be used for classification, regression and also for outlier detection. This algorithm is very effective in high dimensional spaces and where the number of dimensions is greater than the number of given samples. It only uses a small sample of

training points to construct a decision hyperplane which makes it memory efficient. It uses kernel trick for mapping higher dimension space to lower dimension without using actual vectors.

3.3.3 Gradient Boost Classifier

Gradient Boost is a classifier which uses a loss function, weak learners and an additive component to reduce that loss[7]. The loss function depends on the problem statement but the necessary condition is that it should be differentiable. For weak learners, decision trees are used in gradient boosting classifiers. Each tree is added without affecting the other trees and the loss is computed using gradient descent algorithm. The classifier continues to train until and unless a specified number of trees are added or a decent accuracy is achieved. This algorithm is quite powerful but is prone to overfitting. For the algorithm to work correctly it is important that each individual tree remains weak. We achieve this by adding constraints while constructing trees such as tree depth, number of leaves and number of observations per split.

3.3.4 Random Forest Classifier

It is an ensemble learning method which uses multiple trees to give a result[6]. The idea of random forest classifiers is based on the fact that a group of classifiers will outperform a single classifier. Each tree gives its own classification result and the output which occurs most becomes the output of the result. This works very well because individual decision trees are prone to overfitting but multiple trees which are not related to each other corrects each other's mistakes due to which the result produced is more refined. The accuracy of this algorithm depends on the fact that how many different trees are uncorrelated to each other. This is ensured by bagging and feature randomness.

3.3.5 Decision Tree

A decision tree is a flowchart like structure where nodes represent an if else condition, each branch represents the outcome and leaf nodes represent the actual class label[5]. Here, the path from root to leaf gives us the classification rule for that class. Commonly used algorithms for splitting are: Gini impurity, Chi-Square and Information Gain.

3.3.6 Naive Bayes

Naive Bayes is a simple classifier based on Bayes Theorem with a naive assumption that features are independent of each other[10]. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n ,:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{4}$$

Using Naive assumption we get,

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$
$$\Downarrow \quad (5)$$
$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Naive Bayes classifier is quite fast than more complex methods but is a bad estimator due to which its probability values can't be trusted.

3.3.7 Bagging Classifier

It is a meta estimator which makes uses of individual classifiers such as decision trees on a random subset of the dataset, then uses that results to give a final result [9]. It reduces variance by making the construction process randomized. It has been proven that bagging is efficient for unstable algorithms, eg: neural networks and decision trees. This algorithm is closely related to other ensemble algorithms like pasting, random sub spaces and random patches. Its hyper parameters include the number of trees and number of samples. The number of samples and whether replacement is used or not is used to regulate the algorithm.

4 Materials and Methods

4.1 Dataset description

In this work we use two datasets: major ions dataset from US Geological Survey[18] and LUCAS Topsoil dataset[19] from European Commision.

4.1.1 Major Ions Datasets

The data is collected for the US Geological Sur[18] in support of the National Water Census. The dataset contains approximately 66 thousand rows. It contains values of various ions, pH and Electrical Conductivity out of which we only discuss those parameters that are required by us to develop our model for Irrigation Water Quality Index(IWQI) which are: EC, Cl^-, Na^+, HCO^{3-}, Ca^{2+} and Mg^{2+}. EC stands for electrical conductivity which mainly represents the salinity of soil, it was measured in milliSiemens/Centimeter. Cl^-, Na^+, HCO^{3-}, Ca^{2+}, Mg^{2+} represents chloride ion concentration, sodium ion concentration, Bicarbonate ion concentration, Calcium ion concentration and Magnesium ion concentration respectively. All of the ions were measured in milligrams/Litre. They were required to be converted to millimoles/Litres so they were divided by their molar masses. EC needs to be in deciSiemens/Metre

so it is divided by 100. Na^+, Ca^{2+} and Mg^{2+} are used to calculate Sodium Absorption Ratio(SAR) using Equation 1 as mentioned in the previous section. The parameters can also be measured using IoT sensors which is also one of the reasons for choosing them.

4.1.2 Lucas Topsoil Dataset

Lucas dataset was constructed by ESDAC(European Soil Data Centre) to derive statistics for impact of land management on soil properties[19]. We use 2015 lucas topsoil data for our model which has 21859 data points. The dataset consists of physical properties of soil which are clay, silt, sand, coarse fragments and chemical properties of soil which are pH(CaCl2), pH(H2O), Electrical conductivity, Organic carbon content, Carbonates content, Phosphorus content, Total nitrogen content, Extractable potassium content out of which we use pH(H20) and Electrical Conductivity for our model. Electrical Conductivity was given in mS/m which is converted to dS/m for use. pH(H2O) represents Hydrogen Potential measured in water which represents the degree of acidity or basicity of a sample. We chose the above mentioned parameters since it can be measured from an IoT sensor. This dataset also consists of information on land cover and land use, irrigation management, structural elements in the landscape and crop residues which can be used for various applications.

4.2 Irrigation Water Quality Index based classification

This work focuses on classifying various water samples under water quality classes as described in Table 2. We develop a classification model and apply it on the Major Ions Dataset that contains all the parameters necessary to calculate the Irrigation Water Quality Index(IWQI). The IWQI classifier is developed using three steps: Converting parameters to quality measurement values, calculating IWQI using quality measurement values and relative weights of each parameter and assign a class based on ranges in table 2, using various classification techniques to predict IWQI Class and selecting the best technique out of them. These steps are explained in the following subsections.

4.2.1 Quality Measurement Values

We need to normalize the values of parameters in the water quality dataset due to the following reasons: For calculating IWQI, we need 5 parameters which are EC, Na^+, Cl^-, HCO^{3-} and SAR which have different units and hence we need them under a common unit, the value of any parameter being too high or too low will decrease the water quality. Hence, we need to convert the measured values of each parameter according to a set of predefined limits. We convert the value of parameters to the quality measurement value(q_i) that can be obtained using Equation 2 and Table 3. The following Table 5 shows the number of values of each parameter in each range for the major ions

q_i	Q_HCO^{3-}	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 5: Number of Quality measurement values in each range

dataset. For Bicarbonate, we can see that the majority of samples have quality measurement values between 35-85 which won't majorly decrease the quality of water. Electrical Conductivity has the majority of samples having q_i in the range 85-100 which will lead to increase in water quality. Cl^-, Na^+ and SAR have most of their q_i values in the range 0-35 which would decrease the water quality of samples.

4.2.2 IWQI Formula and Interpretation

Irrigation Water Quality Index(IWQI) developed by Meirels et al. is mainly based on salinity and toxicity effects of water samples on soil and plants. It is given by Equation 3 in which we multiply q_i of each parameter with its corresponding weight and sum all the values. The IWQI so obtained is a value between 0-100. Water samples are further separated into classes which are mentioned in Table 2. Water samples having IWQI in the range 85-100 can be used on most of the soils and plants. For Water samples having IWQI in the range 70-85 soils with moderate permeability is preferred and plants with low salt tolerance should be avoided. For IWQI range 55-70, plants with moderate salt tolerance and soils with medium to high permeability is prefered. For samples having IWQI between 40-55, soils with high permeability and plants with moderate to high salt tolerance is preferred along with proper soil control practices. Samples with IWQI less than 40 should not be used for irrigation purposes, There was no proper mention of the salt tolerance limit for plants so we took the limits as defined by salt tolerance[20]. Salt tolerance values for plants is defined by the maximum electrical conductivity of a solution a plant can bear before salinity problem occurs. Plants with salt tolerance less than or equal to 2 dS/m have very low tolerance limit, plants with salt tolerance between 2-4 dS/m have low tolerance limit, plants with salt tolerance between 4-6 dS/m have moderate salt tolerance limit, plants with salt tolerance between 6-8 dS/m have high tolerance limit and plants with tolerance value greater than 8 dS/m have very high tolerance limit. We use the ranges defined above to develop a model for suggesting plants according to water quality class described in section 4.3.

4.2.3 Classification based on IWQI

Water samples can be classified according to IWQI value obtained using Table 2 as explained in the Section 3.1. We can also predict the IWQI class using

a fewer number of parameters than originally in used IWQI calculation using various classification algorithms. Table 6 shows the correlation matrix for various parameters for water samples. Pearson's correlation was used for obtaining the correlation matrix which is obtained using Equation 6. From Table 6 we can infer the following:

1. Cl^- is highly correlated with Na^+ and EC.
2. Na^+ is highly correlated with all the parameters.
3. EC is highly correlated with Cl^-, Na^+ and HCO^{3-}.
4. HCO^{3-} is highly correlated with EC.
5. IWQI is highly correlated with Cl^-, Na^+ and EC.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}} \tag{6}$$

Here,
r: correlation coefficient
X_i: value of x-variable in the sample
$\bar{X}$: mean value of x-variable
Y_i: value of y-variable in the sample
$\bar{X}$: mean value of y-variable

	Cl^-	Na^+	EC	HCO^{3-}	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^{3-}	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 6: correlation matrix for various parameters for water samples

We select three parameters Cl^-, Na^+ and EC for developing our classification model due to them having higher correlation with IWQI. This way, we are able to save cost for sensors for HCO^{3-}, Mg^{2+} and Ca^{2+}(used to calculate SAR). We used 7 classification algorithms for predicting IWQI class for water samples: Decision Tree Classifier, Naive Bayes, Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Bagging and MLP Classifier(ANN Classifier) details of which are mentioned in Section 3.3. The method having the best accuracy and F2 score is selected for further use. Steps for classification are mentioned in algorithm 1.

Algorithm 1 Classification of water samples based on IWQI

Input: Water quality dataset with 5 parameters Cl^-, Na^+, EC, HCO^{3-} and SAR
Output: algorithm with best score
1: Calculate quality measurement values(q_{param}) for each parameter.
2: Calculate IWQI using the Equation 3 and assign classes from Table 2
3: Choose features of dataset as: features $\leftarrow Cl^-$,Na^+,EC
4: Choose labels of dataset as:IWQI
5: Split dataset into training_set and testing_set
6: Apply the 7 classifiers(Decision Tree Classifier, Naive Bayes, Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Bagging and MLP Classifier) on the training set and evaluate on the testing set using accuracy and F2 score.
7: **return** algorithm with best score

4.3 Model for suggesting plants based on EC and pH of soil or water

For suggesting the plants which can be grown in the given water or soil sample, we use the ranges presented in Table 4 to construct an IF - ELSE model which classifies a plant to the given water sample or soil sample. Since soil sample is obtained from lab testing we use data provided by [19] directly but since water dataset can be obtained from IoT sensors, we create a model as described in Section 4.2 to calculate IWQI to find which water is suitable for farming or not. After separating the non-viable samples we use the constructed IF - ELSE model to predict plants for a given water sample. The workflow is indicated by Fig. 1 where electrical conductivity and pH is taken as input for a given water or soil sample and we decide according to the ranges in Table 4 whether water or soil conditions are suitable for the given plant or not. If it meets the required conditions, that water or soil sample is added to the set of water and soil samples for the chosen plant.

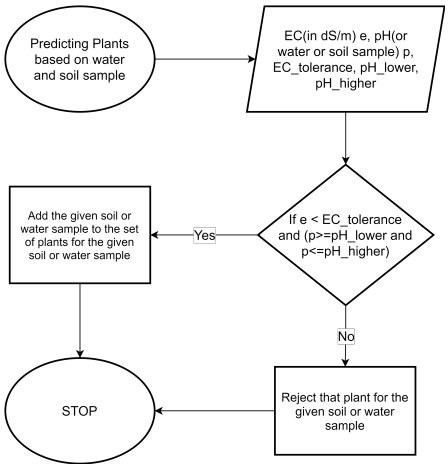


Fig. 1: Calculating IWQI and suggesting the best plants for a given water or soil sample

5 Experiments and Results

The algorithm was implemented in python 3 with pandas and numpy for loading and preprocessing datasets. All of the classification methods were applied using Scikit-learn[23]. The parameter settings for various classification algorithms are described below in 5.1. The evaluation criteria is explained in 5.2 and results in 5.3 and 5.4.

5.1 Parameter Settings for Classification Model

The parameter settings for all the classifiers are mentioned in Table 7. Various parameters for the classifiers are as follows:

- **Decision Tree Classifier:** criterion: The function to measure the quality of a split, splitter: The strategy used to choose the split at each node, min_samples_split: The minimum number of samples required to split an internal node.
- **Naive Bayes Classifier:** alpha: Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing), class_prior:Prior probabilities of the classes, class_count: Number of samples encountered for each class during fitting.
- **Gradient Boosting Classifier:** loss: The loss function to be optimized. learning rate: Learning rate shrinks the contribution of each tree by learning_rate. N_estimators: The number of boosting stages to perform.
- **Random Forest Classifier:** n_estimators: The number of trees in the forest, criterion:The function to measure the quality of a split, base_estimator_: The child estimator template used to create the collection of fitted sub-estimators.
- **Support Vector Classifier:** C: Regularization Parameter, decision_function_shape: The shape of decision function to return, tol:Tolerance for stopping criterion
- **Bagging Classifier:** The base estimator to fit on random subsets of the dataset, n_estimators: The number of base estimators in the ensemble.
- **MLPClassifier:** hidden_layer_sizes: The ith element represents the number of neurons in the ith hidden layer, activation: Activation function for the hidden layer, Learning_rate: Learning rate for updating weights.

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 7: Parameter Settings for Classifiers

5.2 Evaluation Metrics For Classification Model

Classification algorithms used are evaluated using Accuracy, Precision, Recall and F1 score. We also divided the dataset into five parts for cross validation and measured accuracy on each part. Before we define these metrics we need to define True Positive, True Negative, False Positive, False Negative. When a model correctly predicts positive class it is True Positive(TP), correct prediction of Negative class is True Negative(TN), wrong prediction of positive class is False Positive(FP) and wrong prediction of negative class is False Negative(FN). For multi-class classification, true-positive prediction means that an observation has been assigned a class to which it belongs and negatives occur when observation is assigned to any other class. The metrics we used are explained below:

- **Accuracy:** Is defined as the total number of current predictions divided by total number of predictions. Its formula is given by Equation 7.
- **Precision:** Precision evaluates the fraction of correct positives which is given by Equation 8.
- **Recall:** Evaluates the fraction of positives that were identified correctly and is given by Equation 9.
- **F1 Score:** Combines both precision and recall and can be used independently as an evaluation method for classification algorithms. It is given by Equation 10.
- **k-Fold cross validation:** Data is first shuffled and then split into k parts. Each part is taken as a test set in turn while the rest of the k-1 parts are a part of the training set. The model is evaluated using this strategy and we can obtain scores for each of the k parts.
- **Confusion Matrix:** It is a matrix of dimension NXN where N is the number of classes which a classification model predicts. Each row corresponds to the actual prediction for that class and column corresponds to predicted prediction for that class. Each cell in this matrix denotes the number of samples predicted as class X where actual class was Y. This helps us to compute accuracy, precision, recall and F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

5.3 Results of Classification based on IWQI

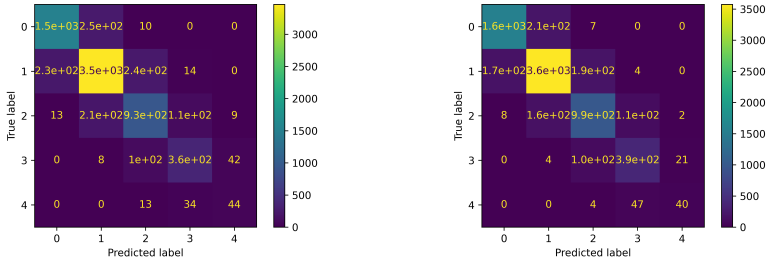
We applied Algorithm 1 on the major ions dataset with 70-30 train-test split and evaluated it with the metrics Accuracy, Precision, Recall and F1 score. The confusion matrix for all the classification algorithms are shown in Fig. 2. For this experiment, we have assigned the numbers to IWQI range with 0-40 as 0, 40-55 as 1, 55-70 as 2, 70-85 as 3 and 85-100 as 4. For the major ions dataset, there are the least number of samples belonging to class 4(495 out of 37999) which is not enough size for classification and hence the poor performance of all the algorithms for that class. For all other classes 0-3, the algorithm performs well except for Naive Bayes Classifier. Using the confusion matrix from Fig. 2 and Equations 7,8,9 and 10, we calculated the Accuracy, Precision, Recall and F1 Score in Table 8. Random forest has the best performance with accuracy of 86.9% followed by Gradient Boosting(85.8%) and MLP classifier(84.6%). Performance of Naive Bayes is the worst with an accuracy of 52.6% and F1 Score of 13.8%. We also performed 5-fold validation for our methods whose results are given in Table 9 where Random Forest performs the best and Naive Bayes performs the worst.

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 8: Classification results for algorithm 1

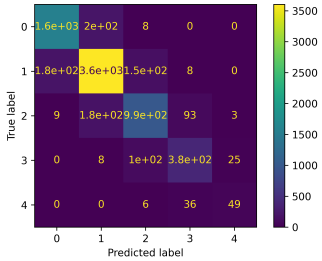
Five Fold Cross Validation					
Methods	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Bagging Classifier	0.829	0.823	0.814	0.830	0.813
DecisionTree	0.833	0.836	0.828	0.823	0.839
Naive Bayes	0.518	0.518	0.518	0.518	0.519
Gradient Boosting	0.862	0.866	0.869	0.859	0.862
Random Forest	**0.866**	**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 9: Five fold cross validation scores for algorithm 1

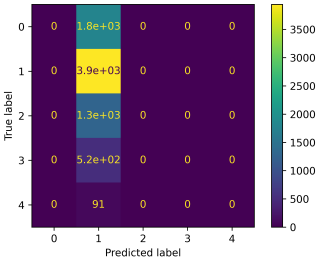


(a)

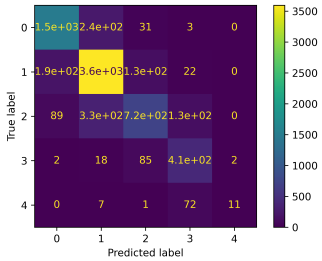
(b)



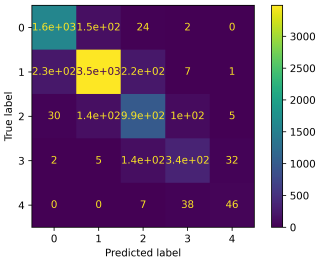
(c)



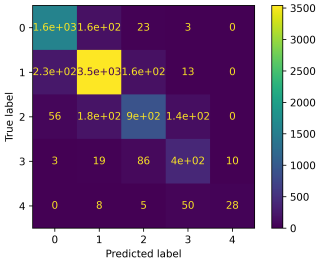
(d)



(e)



(f)



(g)

Fig. 2: Confusion matrices for (a)Decision Tree, (b)Gradient Boost, (c)Random Forest, (d)Naive Bayes, (e)Bagging Classifier, (f)ANN, (g)SVC

5.4 Results of model for suggesting plant based on a given water or soil sample

We apply the model given in Section 4.3 for suggesting plants based on a given water or soil sample. We suggest plants for 21,859 soil samples from lucas dataset[19] and 37,999 water samples from major ions dataset[18]. Some soil and water samples didn't have suitable conditions so they were rejected. Plants were suggested for 13,700 soil samples and 29,114 water samples. 8,159 soil and 8,885 water samples were rejected. Results of 20 soil and water samples are given in Table 10 and Table 11 respectively.

Soil_sample	**List of plants suitable for the given soil sample**
35163814	'Potatoes', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax'
34463934	'Safflower'
33983238	'Safflower'
34043240	'Potatoes', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax'
33723266	'Safflower'
34123260	'Safflower', 'Fall rye', 'Oats'
33523262	'Safflower'
33683636	'Corn', 'Barley', 'Wheat ', 'Carrots', 'Strawberry', 'Tomato', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
34983634	'Safflower'
33623332	'Garden Beet', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax'
34383316	'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax'
34083328	'Safflower', 'Fall rye', 'Oats'
34543252	'Garden Beet', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax'
34883940	'Safflower'
34683874	'Safflower', 'Fall rye', 'Oats'
34823880	'Safflower', 'Fall rye', 'Oats'
35623878	'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
36143890	'Corn', 'Barley', 'Wheat ', 'Tomato', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
35523886	'Safflower'
35343868	'Safflower'
34903806	'Safflower', 'Fall rye'

Table 10: List of Plants for 20 soil samples out of 13,799 samples

Water_sample	List of plants suitable for the given water sample
0	'Garden Beet', 'Potatoes', 'Corn', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
1	'Garden Beet', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
2	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
3	'Garden Beet', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
51	'Garden Beet', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
52	'Garden Beet', 'Potatoes', 'Corn', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
53	'Garden Beet', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
54	'Garden Beet', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
55	'Garden Beet', 'Potatoes', 'Corn', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
4	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
5	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
6	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
7	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
8	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
9	'Garden Beet', 'Potatoes', 'Corn', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
10	'Garden Beet', 'Potatoes', 'Corn', 'Barley', 'Wheat ', 'Cabbage', 'Tomato', 'Broccoli', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye', 'Oats', 'Yellow Mustard', 'Flax', 'Canola'
11	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
12	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
13	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'
14	'Garden Beet', 'Barley', 'Wheat ', 'Asparagus', 'Spinach', 'Sunflower', 'Kochia', 'Sugar beet', 'Safflower', 'Fall rye'

Table 11: List of Plants for 20 water samples out of 29,114 samples

6 Conclusion and Future Work

The purpose of this work was to develop a classification model for classifying water samples based on irrigation quality and to suggest plants based on given water quality and soil salinity and pH. The work can be used to save the cost of lab tests involved in measuring water and soil samples once we start using sensors for measuring the properties. For quantifying water quality, we calculated Irrigation Water Quality Index(IWQI) of water samples using properties like Electrical Conductivity and Chemical Ion concentrations like Sodium, Chlorine etc of water. The index chosen is mainly focused on the effect of water on soil sodicity, salinity and toxicity. We assigned water quality classes to different ranges of IWQI values. Seven classification methods were used out of which Random Forest performed the best(86.9%) followed by Gradient Boosting and Neural Networks on all of the evaluation methods used. For suggestions of plants based on water and soil properties, we converted the restrictions on soil salinity and pH and water quality for various plants to conditional statements. We were able to suggest plants for all the soil and water samples obtained except one third of soil and one fifth of water samples due to them being unfit for agricultural purposes.

In future, we plan to extend our model by automating data collection processes using IoT devices. We plan to develop a recommendation system of crops based on soil, water and weather properties. We plan to work on other phases of agriculture such as soil preparation, crop selection, fertilizing and harvesting and developing models to analyze and assist in them. We plan to develop a mobile/web application that will provide an interface for using functionalities developed in this work. We also plan to explore other possible domains of IoT such as networking, developing softwares for hardware like raspberry PI and Arduino and automation in Farming.

References

1. S. Sunder, "India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more r&d needed," *The Financial Express*, 2018.
2. N. Zhang, M. Wang, and N. Wang, "Precision agriculture—a worldwide overview," *Computers and electronics in agriculture*, vol. 36, no. 2-3, pp. 113–132, 2002.
3. D. La Mora-Orozco, H. Flores-Lopez, H. Rubio-Arias, A. Chavez-Duran, J. Ochoa-Rivero *et al.*, "Developing a water quality index (wqi) for an irrigation dam," *International journal of environmental research and public health*, vol. 14, no. 5, p. 439, 2017.
4. J. L. Lerios and M. V. Villarica, "Pattern extraction of water quality prediction using machine learning algorithms of water reservoir," *International Journal of Mechanical Engineering and Robotics Research*, vol. 8, no. 6, 2019.
5. Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.
6. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
7. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
8. M. H. Hassoun *et al.*, *Fundamentals of artificial neural networks*. MIT press, 1995.
9. L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

10. K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, p. 60, 2006.
11. A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "Iot and machine learning approaches for automation of farm irrigation system," *Procedia Computer Science*, vol. 167, pp. 1250–1257, 2020.
12. B. Gupta and B. Huang, "Mechanism of salinity tolerance in plants: physiological, biochemical, and molecular characterization," *International journal of genomics*, vol. 2014, 2014.
13. P. Shrivastava and R. Kumar, "Soil salinity: a serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation," *Saudi journal of biological sciences*, vol. 22, no. 2, pp. 123–131, 2015.
14. R. Gentili, R. Ambrosini, C. Montagnani, S. Caronni, and S. Citterio, "Effect of soil ph on the growth, reproductive investment and pollen allergenicity of ambrosia artemisiifolia l." *Frontiers in plant science*, vol. 9, p. 1335, 2018.
15. M. E. Grismer and K. M. Bali, "Drought tip: Use of saline drain water for crop production," 2015.
16. A. C. M. Meireles, E. M. d. Andrade, L. C. G. Chaves, H. Frischkorn, and L. A. Crisostomo, "A new proposal of the classification of irrigation water," *Revista Ciencia Agronomica*, vol. 41, no. 3, pp. 349–357, 2010.
17. R. S. Ayers, D. W. Westcot *et al.*, *Water quality for agriculture*. Food and Agriculture Organization of the United Nations Rome, 1985, vol. 29.
18. S. Qi and A. Harris, "Geochemical database for the brackish groundwater assessment of the united states," *US Geological Survey data release*, 2017.
19. A. Orgiazzi, C. Ballabio, P. Panagos, A. Jones, and O. Fernández-Ugalde, "Lucas soil, the largest expandable soil dataset for europe: a review," *European Journal of Soil Science*, vol. 69, no. 1, pp. 140–153, 2018.
20. V. Chinnusamy, A. Jagendorf, and J.-K. Zhu, "Understanding and improving salt tolerance in plants," *Crop science*, vol. 45, no. 2, pp. 437–448, 2005.
21. R. Gentili, R. Ambrosini, C. Montagnani, S. Caronni, and S. Citterio, "Effect of soil ph on the growth, reproductive investment and pollen allergenicity of ambrosia artemisiifolia l." *Frontiers in Plant Science*, vol. 9, p. 1335, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpls.2018.01335
22. W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
23. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.