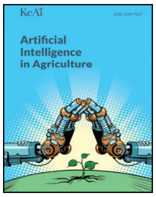
Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Nutrient optimization for plant growth in Aquaponic irrigation using Machine Learning for small training datasets

Sambandh Bhusan Dhal [a,*], Muthukumar Bagavathiannan [b], Ulisses Braga-Neto [a], Stavros Kalafatis [a]

[a] Department of Electrical and Computer Engineering, Texas A&M University, USA
[b] Department of Soil and Crop Sciences, Texas A&M University, USA

ARTICLE INFO

Article history:
Received 14 January 2022
Received in revised form 18 May 2022
Accepted 19 May 2022
Available online 24 May 2022

Keywords:
Hydroponic
Aquaponic
Training datasets
Non-linear algorithms
Semi-bolstered error estimation
Linear support vector machine
Decision Support System

ABSTRACT

With the recent trends in urban agriculture and climate change, there is an emerging need for alternative plant culture techniques where dependence on soil can be eliminated. Hydroponic and aquaponic growth techniques have proven to be viable alternatives, but the lack of efficient and optimal practices for irrigation and nutrient supply limits its applications on a large-scale commercial basis. The main purpose of this research was to develop statistical methods and Machine Learning algorithms to regulate nutrient concentrations in aquaponic irrigation water based on plant needs, for achieving optimal plant growth and promoting broader adoption of aquaponic culture on a commercial scale. One of the key challenges to developing these algorithms is the sparsity of data which requires the use of Bolstered error estimation approaches. In this paper, several linear and non-linear algorithms trained on relatively small datasets using Bolstered error estimation techniques were evaluated, for selecting the best method in making decisions regarding the regulation of nutrients in hydroponic environments. After repeated tests on the dataset, it was decided that Semi-Bolstered Resubstitution Error estimation technique works best in our case using Linear Support Vector Machine as the classifier with the value of penalty parameter set to one. A set of recommended rules have been prescribed as a Decision Support System, using the output of the Machine Learning algorithm, which have been tested against the results of the baseline model. Further, the positive impact of the recommended nutrient concentrationson plant growth in aquaponic environments has been elaborately discussed.

1. Introduction

Aquaponics is an agricultural technique that is touted as a promising alternative to solve the food and environmental crisis that the world faces today. As it is a combination of aquaculture (Pillay, 2003; Pillay and Kutty, 2009) and hydroponics (Jensen, 1997; Roberto, 2003; Jones Jr, 2016), this technique can have high water use efficiency, and reduce the dependency on pesticides and fertilizers which adds to the sustainability of the system. This system has been shown to consume 50–70% less water compared to traditional agricultural systems, owing to its recyclability. Moreover, this system requires less pest control and has proven to be less impacted by harsh weather conditions, leading to yield increase.

As we are moving towards the era of digital agriculture, efforts have been made to design physical prototypes for smart commercial aquaponic systems. Lobanov et al. has focused on proposing a new system for reducing the carbon and nitrogen content from fish waste to produce liquid fertilizer which is added to facilitate the growth of lettuce in aquaponic set-ups (Lobanov et al., 2021). Karimanzira et al. introduced the concept of building an intelligent aquaponics system incorporating predictive analysis, system optimization and anomaly detection for maximizing productivity in commercial aquaponics through early fault detection (Karimanzira et al., 2021). Mahanta et al. formulated a laboratory set-up as a prototype to grow soybeans in hydroponic solution using plasma activated water to decrease the amount of heavy metal uptake and optimize yield (Mahanta et al., 2022). Rau et al. designed a smart IoT based sensing and actuation system for growing rice by controlling the concentration of magnesium and nitrogen in hydroponic solution along with monitoring the environmental parameters of the greenhouse (Rau et al., 2017). A similar design was proposed by Dhal et al. using a smart IoT system for real time sensing and regulation of nutrients in commercial aquaponic set-ups depending on the season in which the crops were grown (Dhal et al., 2022). Timsina et al. proposed the use of Machine Learning models to regulate nutrients for growing cereals in farmlands, but efforts are yet to be made for growing them in aquaponic set-ups (Timsina et al., 2021).

* Corresponding author at: Department of Electrical and Computer Engineering, Texas A&M University, USA

E-mail addresses: sambandh@tamu.edu (S.B. Dhal), muthu.bagavathiannan@tamu.edu (M. Bagavathiannan), ulisses@tamu.edu (U. Braga-Neto), skalafatis-tamu@tamu.edu (S. Kalafatis).

https://doi.org/10.1016/j.aiia.2022.05.001

There have been recent advancements in the field of Smart Aquaponics which involve monitoring environmental parameters as well as plant growth through different Machine Vision-Based approaches in an IoT environment (Arvind et al., 2020; Lauguico et al., 2020). Arvind et al. implemented an AutoML model trained with an XGradientboost algorithm, with 10-fold cross-validation taking into account the different sensor values recorded in the greenhouse and the fish count in the aquaponic tank which was extracted using the mask R-CNN image segmentation. This algorithm was used to control the triggering of the actuators in the system that in turn control the environmental parameters in the greenhouse; ensuring significant improvements in yield and water conservation, when compared with the conventional methods. Languico et al. did a comparative study of three ML estimators: K-Nearest Neighbours (KNN), Logistic Regression (LR), and Linear Support Vector Machine (L-SVM), on the vision-feature extracted images of lettuce in a smart aquaponics set-up to monitor diseases that the crop may incur in its lifetime. A similar kind of study was done by Maleki-Kakelar et al. where multiple ML algorithms like linear and quadratic regression models, fuzzy systems and genetic programming was used to conduct regression analysis for improving urease activity aimed at strengthening the behaviour of soils (Maleki-Kakelar et al., 2021). A study on images of lettuce leaves was conducted by Concepcion II et al. to detect diseases, wherein different feature selection processes were used to select the top four attributes for training the ML models (Concepcion II et al., 2020). A study on smart nutrient regularization for replenishing the Nitrogen, Phosphorus and Potassium content of the soil has been done by Ahmed et al. using genetic algorithms to provide the optimal level of nutrients needed for high production level of crops (Ahmed et al., 2021). Hiram Ponce et al. used a combination of Convolutional Neural Network (CNN) for extracting features from tomato leaves along with a combination of Artificial Hydrocarbon Network (AHN) as the dense layer to predict deficiency of nutrients in tomato plants (Ponce et al., 2021). A similar kind of Deep Neural Network was implemented by Yadav et al. for apple foliar disease classification using Plant Pathology image datasets (Yadav et al., 2022). Nevertheless, limited research has been conducted on monitoring and regulating nutrient concentrations in the aquaponic solution using ML-based approaches. The current work focuses on building a ML algorithm that monitors the nutrient status of hydroponic irrigation water and outputs a recommendation system for regulation of these parameters. The main motivation behind this entire approach is to select the most important nutrients that need to be regulated in aquaponic environments depending on the output of Machine Learning classifiers trained on small datasets. The main issues which one may face while designing such a data-driven approach has been discussed in the next paragraph.

One of the major challenges with automation in aquaponics is the lack of sufficient data and the vast number of predictors that have to be used for making any inferences. This could result in what is referred to as the “Curse of Dimensionality” leading to the available data becoming sparse (Köppen, 2000). Thus, it becomes extremely important to reduce the dimensionality of the dataset without losing valuable information. For this, feature selection (Chandrashekar and Sahin, 2014) techniques become more relevant. To state a few, Recursive Feature Elimination (Granitto et al., 2006) and ensemble techniques such as the ExtraTreesClassifier (Shafique et al., 2019) have proven to be highly effective. It is also equally important to check for the separability of the classes in the dataset. Many data visualization techniques like Principal Component Analysis (Wold et al., 1987; Abdi and Williams, 2010) and Multi-Dimensional Scaling (Cox and Cox, 2008) plots can be used to understand how linearly separable the data is and what classifiers would be best suited for the purpose.

Another major drawback especially in the case of small datasets is the problem of “overfitting” the data (Dietterich, 1995). Traditional ML and Deep Learning (DL) algorithms have a high probability of performing poorly on small datasets. A solution to this problem was suggested by Shao et al. who proposed deep Reinforcement Learning algorithms named MONEADD and did a comparative study with Knapsack and Traveling Salesman problem to design neural networks for different combinatorial optimization problems without much feature engineering which showed better scalability when distributed on multiple GPUs (Shao et al., 2021). On a similar note, to address this problem of overfitting, Braga-Neto et al. proposed Bolstered Error Estimation method, which uses the same data for both classifier design and error estimation (Braga-Neto and Dougherty, 2004; Sima et al., 2004). In this form of error estimation, the variance setting for the Bolstering kernel is determined in a non-parametric manner from the data. For all the linear classification rules, the integrals in the Bolstered error estimation are computed in the same way. For the non-linear classification rules, a small number of Monte-Carlo samples are generated. In this study, three types of Bolstered error estimators, namely the Gaussian-Bolstered resubstitution error estimator, semi-Bolstered resubstitution error estimator, and the Gaussian Bolstered Leave-One-Out error estimator, have been used with linear classifiers like LDA and SVC, along with non-linear classifiers like CART and KNN, and their results have been compared to identify the best performing classifier in this case. Based on the performance of the classifier with utmost optimal performance, a set of recommendation rules were prescribed and a comparative study was done on how this proposed Machine Learning based approach resulted in more optimal yield as compared to the baseline model.

2. Methodology

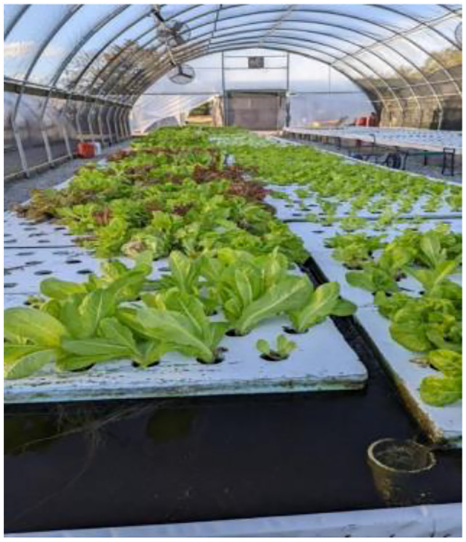
The dataset which was used in this case was recorded from three commercial aquaponic facilities located in Caldwell, Bryan, and Grimes counties in Texas, USA which are large producers of lettuce and other greens. From the aquaponic facility at Caldwell, two water samples were collected per week: one from the fish tank and the other from the plant bed which was used to grow Romaine Lettuce, watercress, lettuce and green peppers. Similarly, from the facility at Bryan, three water samples were collected: one from the Tilapia tank, one from the Goldfish tank, and the other from the main plant bed where lettuce and kale were cultivated. The aquaponic facility at Grimes was one of the largest in the state and four water samples were collected each week: one each from the main growth tank which was used for breeding Tilapia and certain shrimps, one from the tank which bred bluegill fish, one from the tank which was used to grow plant seedlings and the other from the main greenhouse where tomatoes, lettuce and collard greens were grown. All of these tanks were connected together with continuous water flow between them, and water purifiers were installed in each of these tanks to ensure that the recycled water meets the optimal water quality requirements needed for growth of fish and plants. A set up where the experiments were recorded have been shown in Fig. 1.

The wastewater from the fish tank is connected to the main set-up, where extra nutrients are added to the aquaponic solution to optimize plant growth. After collecting the water samples, they were sent to the Soil, Water, and Forage Testing Laboratory, Texas A&M University to determine the nutrient concentration for each of the samples. The method used to carry out each of these nutrient concentrations in the laboratory is described as follows in Table 1.

This process was carried out each week for 9 months and then the data were analysed. Before explaining the analysis part, a pipeline on how the data was processed to design the Decision Support System has been elaborately stated in Fig. 2 below.

2.1. Data analysis

In this study, a total of 143 observations were used with a total of 21 predictors. The initial set included 11 predictors for chemicals namely calcium, magnesium, sodium, potassium, boron, carbonate, bicarbonate, sulfate, chloride, nitrates, and potassium (all of these measured in



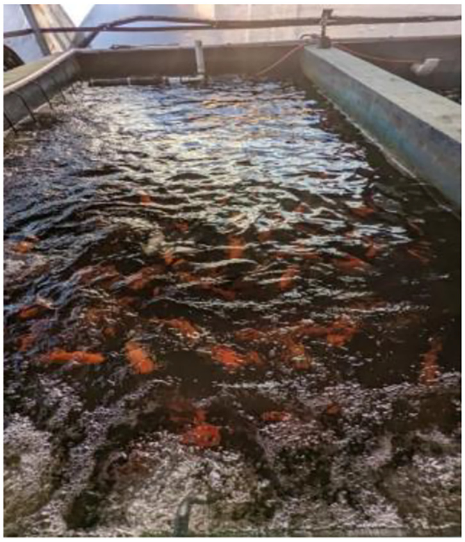


Fig. 1. An aquaponic set-up in aquatics Green Farm, Bryan, TX showing the tank with the lettuce plant bed (Left) and The Goldfish tank which provided nutrients for the plants (Right).

ppm); 8 predictors that define the property of the solution namely the pH, conductivity (measured in umhos/cm), 2 measures of hardness (one measured in grams of calcium carbonate per gallon and the other measured in ppm), alkalinity (measured in ppm of calcium carbonate), Total Dissolved Salts (measured in ppm) and Sodium Adsorption Ratio (SAR) coupled with two categorical predictors (one stores the location of the farm and the other stores the month in which the observations were recorded). The months in which the data were recorded were binned into 5 categories as follows: January to March (category I), April and May (category II), June to August (category III), September and October (category IV), November and December (category V).

To have a clear understanding of the data, the entire dataset was treated with an unsupervised approach. K-means algorithm (algorithm = 'auto', number of iterations = 600) (Likas et al., 2003) was used to classify the datapoints into either of the two classes.

2.2. Dimensionality reduction

As mentioned above, due to the limited size of the dataset, it is not possible to make accurate inferences taking all the predictors into account. Thus, several dimensionality reduction techniques have been used on the dataset to select the top predictors that define the nutrient concentration of the aquaponic solution. The predictors which had zero variance were removed from the dataset. Then, a correlation matrix was constructed between the predictors, and one of the two predictors, which had higher than 90% pair-wise correlation among them, was removed.

Table 1
Method of measurement of different chemical properties in the Aquaponic solution (Provin and Pitt, 2002).

Sl. no.	Property of the aquaponic solution	Process of measurement
1	pH	Using Hydrogen ion-selective electrode
2	Conductivity	Using conductivity probe
3	Nitrate concentration	By reduction of nitrite to nitrate using a cadmium column followed by spectrophotometric measurement
4	Chloride concentration	Chromatography
5	Carbonate/Bicarbonate concentrations, Hardness, and Alkalinity	Carbonate/Bicarbonate conc. Are measured by acid titration using sulphuric acid; water alkalinity is measured from carbonate/bicarbonate concentrations; hardness is calculated from Ca and Mg conc.

Next, all the predictors with less than 5% importance in the dataset were eliminated as they would likely incorrectly skew any inferencing made from the dataset. As the primary goal was to bring down the size of the dataset to 5 primary chemical predictors, Recursive Feature Elimination technique with XGBoost classifier was used, which ranked the predictors in the order of their importance. This resulted in bringing down the size of the dataset to 5 chemical predictors. In addition to this, 2 categorical predictors were also appended to the dataset, one storing the month and the other storing the place in which the observations were recorded. Therefore, a total of 143 observations with 7 predictors were used to design classification rules and carry out inferences.

2.3. Data visualization

To perform any classification rules on a given dataset, Data Visualization is an important tool as it aids in understanding the structure of the data. It equally helps in choosing the classification techniques that can be used on the dataset depending on the separability between the classes. In this case, Principal Component Analysis was used. As PCA treats the entire analysis as an unsupervised learning approach and performs an orthogonal transformation of the data, Principal Components were calculated to visualize the variance in the dataset.

The loading matrix was analysed to determine the predictor that contributes the most in each Principal Component and therefore, an inference can be drawn about the relative importance of each predictor from their holding values depending on the value of correlation, which was observed between the predictors and the respective PCs.

To further explore the interpretation from the PCs, the pairwise PC plots have been studied to infer which classifiers would suit the best depending on the pattern of separability of the data. All the datapoints belonging to class 0 have been color-coded in red and the data points belonging to class 1 have been color-coded in green to have a clear visual understanding of the binary distribution of the data.

2.4. Classification techniques and error estimation algorithms

As stated above, the size of the dataset poses a serious issue while designing the classifier in this context, due to which, the classifiers

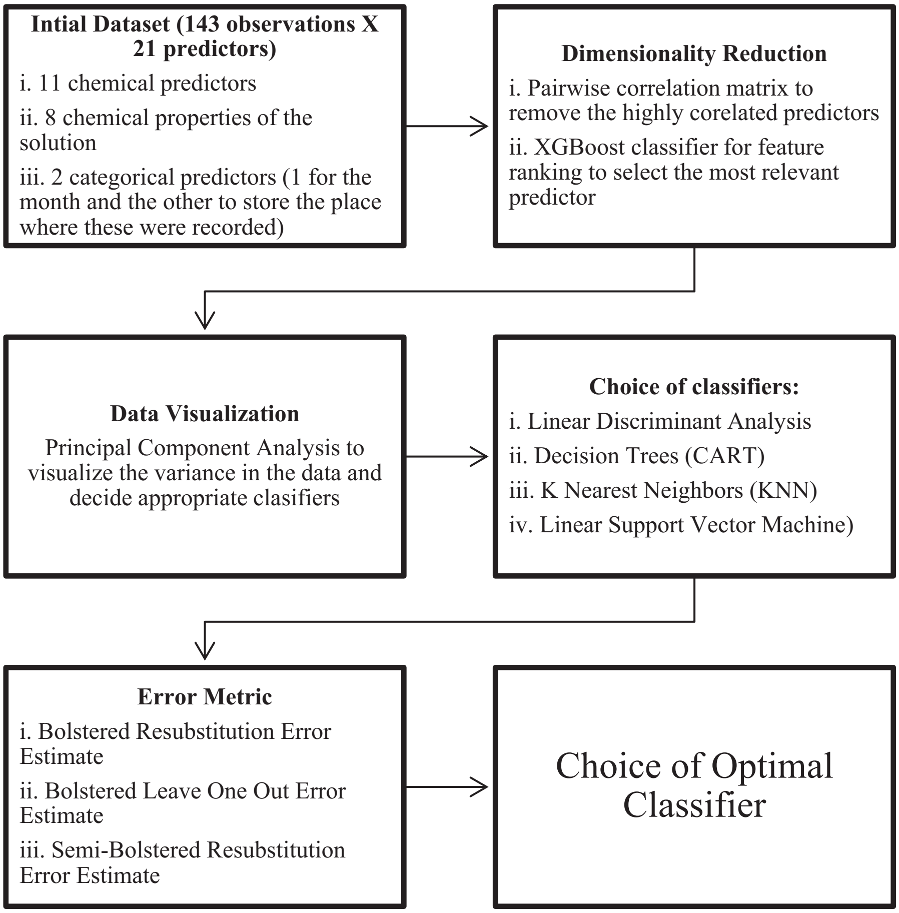

Intial Dataset (143 observations X 21 predictors)
i. 11 chemical predictors
ii. 8 chemical properties of the solution
iii. 2 categorical predictors (1 for the month and the other to store the place where these were recorded)
Dimensionality Reduction
i. Pairwise correlation matrix to remove the highly corelated predictors
ii. XGBoost classifier for feature ranking to select the most relevant predictor
Data Visualization
Principal Component Analysis to visualize the variance in the data and decide appropriate clasifiers
Choice of classifiers:
i. Linear Discriminant Analysis
ii. Decision Trees (CART)
iii. K Nearest Neighbors (KNN)
iv. Linear Support Vector Machine)
Error Metric
i. Bolstered Resubstitution Error Estimate
ii. Bolstered Leave One Out Error Estimate
iii. Semi-Bolstered Resubstitution Error Estimate
Choice of Optimal Classifier

Fig. 2. Schematic of the data processing pipeline to design the Decision Support System for the proposed aquaponic set-up.

have been trained and tested on the same data. This method of error estimation is referred to as Bolstered Error Estimation. One of the main reasons for choosing this type of estimator is its low bias as well as low variance. This also results in a faster estimator compared to other resampling methods like the bootstrap (Hesterberg, 2011). The error estimates have been calculated for each of these estimators for the four popular classification rules by varying the size from small to moderate-sized datasets. The basic idea is to bolster the original empirical distribution of the available data utilizing suitable Bolstering kernels placed at each datapoint location. For this case, a uniform zero-mean, spherical Bolstering kernel f_i° was chosen for analysis, with covariance matrices of the form $\sigma_i^2 I_p$. In each case, there would be a family of Bolstered estimators corresponding to each value of $\sigma_i$where i varies from 1 to n. Larger values of σ_i would result in wider Bolstering kernels resulting in lower variance estimators, but after a point, bias starts to increase. Therefore, choosing the values of standard deviations (SD) for the Bolstering kernels is a challenge and several approaches have been attempted to find the best error estimator in this case which minimizes the bias-variance trade-off.

The value of SD for the Bolstering kernels is given by:

$$\sigma_i = \frac{\hat{d}(y_i)}{\alpha_{p,i}} \; for \; i = 1, 2, \ldots, n \tag{1}$$

here, $\alpha_{p,i}$ is the “constant” correction factor which is a function of the dimensionality of the dataset. $\hat{d}(y_i)$ is the mean minimum distance between the points belonging to the class y_i and is computed using the equation:

$$\hat{d}(y_i) = \left(\sum_{i=1}^{n} \min_{j\neq i} \{||x_i - x_j||\} I_{y_{i=y}}\right) / \left(\sum_{i=1}^{n} I_{y_i=y}\right) \tag{2}$$

Based on the value of calculated standard deviations of these Bolstered kernels, three types of Bolstered Error estimators namely the Bolstered Resubstitution error, Bolstered Leave-One-Out error, and semi-Bolstered Resubstitution error estimates were calculated in conjunction with linear (LDA and Linear SVM) and non-linear classifiers (KNN and CART), and depending on the results, an appropriate decision was made as to which classifier would suit our purpose.

3. Results

As stated before, the entire approach was treated as an unsupervised approach and the K-means algorithm (algorithm = ‘auto’, number of iterations = 600) was used to classify the observations into two classes. Out of the 143 observations used for analysis, 84 were classified into class 0 and 59 into class 1. As it was difficult to derive inferences considering all the predictors, dimensionality reduction techniques were used. ‘Carbonate (ppm)’ was dropped as the variance was zero throughout. Then, a correlation matrix (Fig. 3) was constructed between the rest of the predictors.

This led to the removal of 6 predictors from the dataset namely magnesium, both the measures of hardness, alkalinity, Total Dissolved Salts, and conductivity. Then, ExtraTreesClassifier (Fig. 4) was used to find the percentages of importance for each of the 14 predictors.

Now, the predictors that had less than 5% importance were removed from the dataset. From the list of chemical predictors, Nitrates and

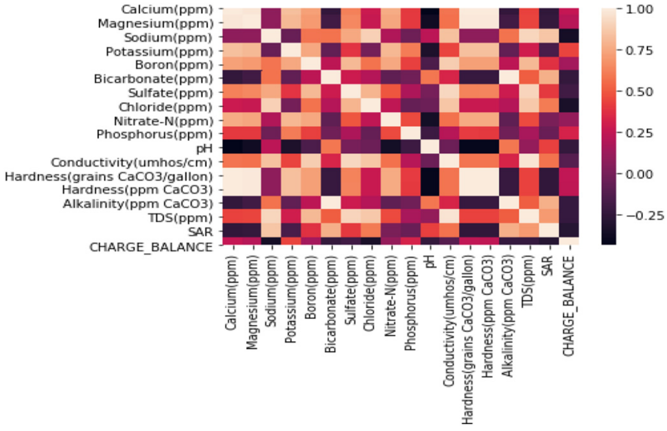


Fig. 3. Color-coded correlation matrix which shows the percentage of correlation between the predictors.

Phosphorus, with the importance of 4.51% and 4.44% respectively, and from the list of chemical properties, pH, SAR, and charge balance with the importance of 1.2%, 0.93%, and 0.93% respectively were removed. Thus, the final list of chemical predictors used in the analysis was as follows: Potassium, Boron, Bicarbonate, Sulfate, and Chloride concentrations in the solution (all of these measured in ppm). In addition to this, 2 categorical predictors were also appended to the dataset, one storing the month and the other storing the place in which the observations were recorded. Therefore, a total of 143 observations with 7 predictors were used to design classification rules and to carry out inferences.

Next, visualizations were carried out on the dataset to gauge the separability of the data and the classifiers which can be used for inferencing. Principal Component Analysis was carried out on the dataset to visualize the dataset in orthogonal projections.

As inferred from Fig. 5, the first three PCs go on to explain 46.34%, 34.40%, and 9.36% (total 90.1%) of the total variance in the dataset respectively. The loading matrix was analysed to get an idea about which predictor contributes the most in each PC and therefore, an inference can be drawn about the relative importance of each predictor from their holding values.

From Table 2, it can be inferred that the PLACE_CLASS, MONTH_CLASS, and the bicarbonate variables had the highest holding values in the 1st, 2nd, and 3rd PC respectively, indicating their strong correlations with the respective PCs. From this, one of the strongest inferences that can be drawn is that both the categorical predictors, one storing the place and the other storing the month in which the observations were recorded, are the most important predictors as they show high holding values in the first two PCs which explain about 80% of the variance in the dataset.

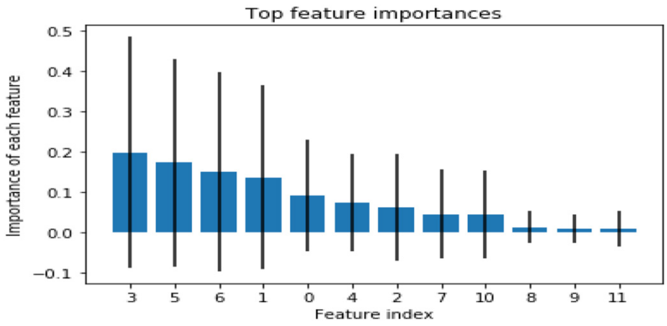


Fig. 4. Percentages of Importance for each of the predictors as identified by ExtraTreesClassifier.

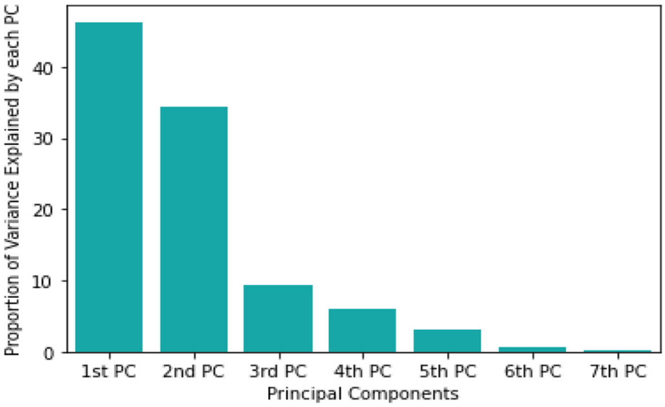


Fig. 5. The proportion of the variance in the dataset explained by each PC.

To further explore the interpretation from the PCs, the pairwise PC plots (Fig. 6) have been studied to infer which classifiers would suit the best depending on the pattern of separability of the data. All the datapoints belonging to class 0 have been color-coded in red and those belonging to class 1 in green, giving a clear visual understanding that the data follows a binary distribution.

From the PC plots, it has been observed that most of the discriminatory information is contained in the first PC, as expected, as it explains most of the variance in the dataset.

Next, due to the low sample size, there is a high probability of overfitting the data because of which, the notion of different Bolstered error estimators has been introduced in this domain. The performance of linear classifiers like LDA and Linear SVM along with the performance of non-linear classifiers like CART (Depth = 2) and 3-NN have been used along with these Bolstered error estimators, and results have been discussed below. The value of the depth of the decision tree was chosen as 2 due to the low volume of the dataset as a higher value of the depth would result in overfitting the data and lead to unreasonably high accuracy on the training dataset. The value of the K for the K-Nearest Neighbor was chosen to be 3 using Elbow method as the mean cluster distance on the training dataset was optimal.

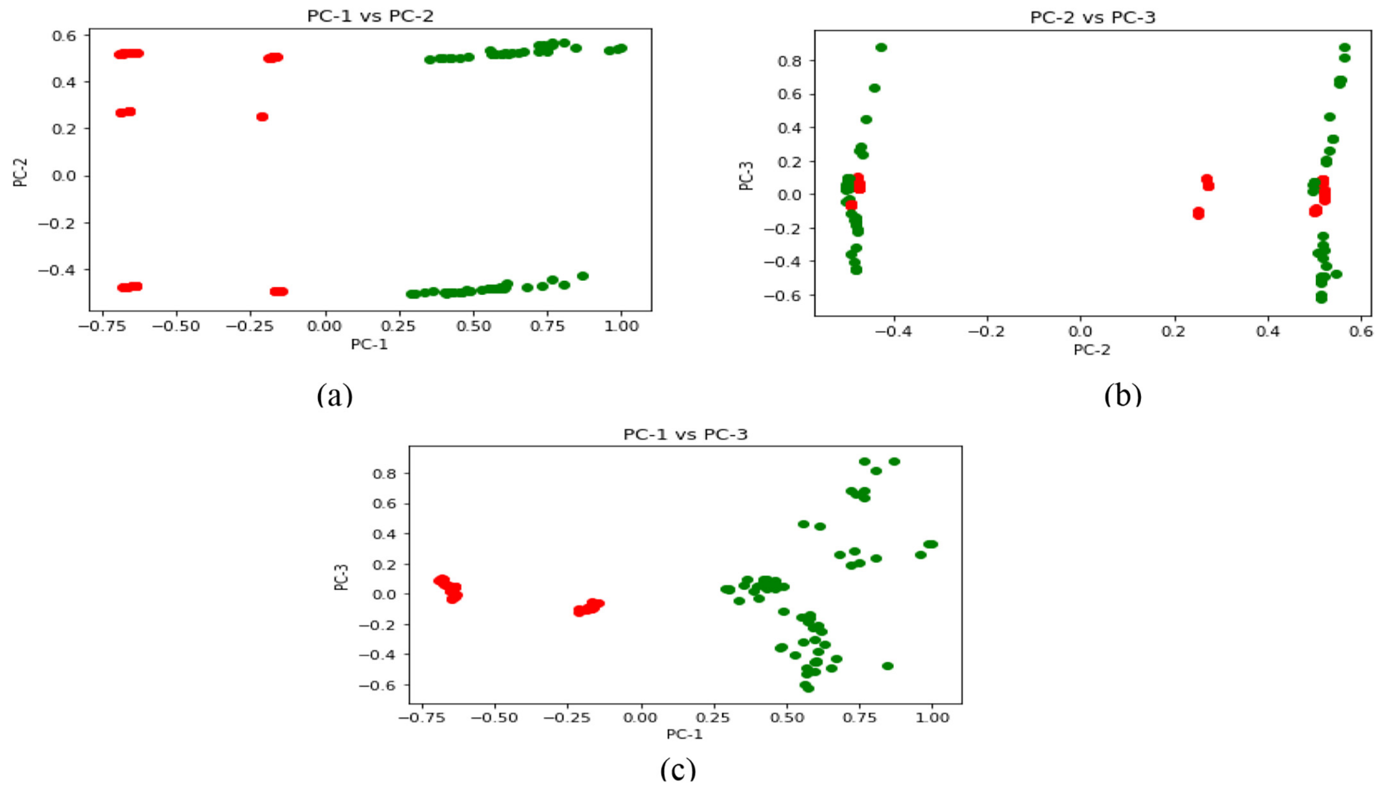
3.1. Bolstered resubstitution error estimation

This type of error is calculated as the sum of all error contributions divided by the total number of points. For a linear classifier where the decision boundary is a hyperplane, the analytical expressions for the integrals were calculated. For the non-linear classifiers, a small number of Monte-Carlo samples were generated (M = 10). The expression for a general Bolstered resubstitution estimator is as follows:

$$\epsilon^{\circ}_{resub} = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{A_1} f_i^{\circ}(x-x_i)dxI_{y_i=0} + \int_{A_0} f_i^{\circ}(x-x_i)dxI_{y_i=1}\right) \qquad (3)$$

Table 2
Loading matrix for each principal component.

Sl. no.	Name of the predictor	Holding values for 1st PC	Holding values for 2nd PC	Holding values for 3rd PC
1	Potassium	0.14	0.05	0.52
2	Boron	0.36	0.038	0.39
3	Bicarbonate	0.20	0.021	**−0.73**
4	Sulfate	0.37	0.0006	0.11
5	Chloride	0.27	0.03	−0.03
6	PLACE_CLASS	**0.77**	−0.042	−0.13
7	MONTH_CLASS	−0.002	**0.99**	−0.03



(a)

(b)

(c)

Fig. 6. (a), (b) and (c). Pairwise PC plots showing separability between the data points belonging to classes 0 and 1.

Eq. (3) holds in the case of linear classifiers like LDA and linear SVM where the Bolstering kernels are given by uniform circular distributions. As the decision boundary in the above case is a hyperplane, it is possible to find analytical solutions as in (1). However, when the error estimate is made for non-linear classifiers like CART and KNN, an approximate solution is needed which is obtained by applying Monte-Carlo integration.

The error estimates for all the four classifiers i.e. LDA, KNN($K = 3$), CART and Linear SVM have been tabulated and the results are shown below. The value of N i.e. the number of randomly sampled datapoints has been varied from 20% to 100% in increments of 10 for each of the classifiers.

As shown in Fig. 7, the Bolstered Error Estimates reduced from 8% to 2.5% for LDA when the size of the sample set increased. For the non-linear classifiers, it can be observed that they overfit the training set due to their minuscule size thereby increasing the chance of the classifier performing badly on the testing set. Therefore, a decision not to proceed with either of these classifiers was made using Bolstered Error resubstitution as it is very likely to overfit when the size of the dataset is small.

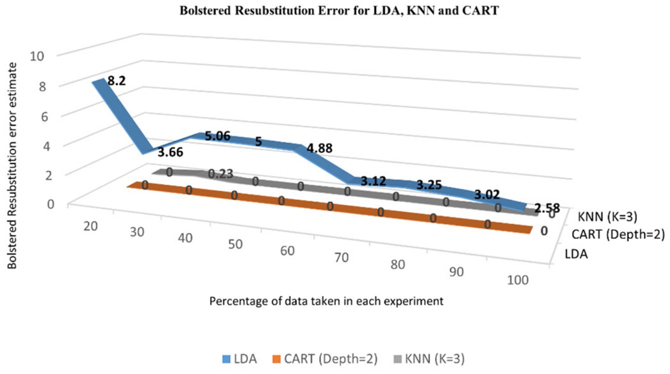


Fig. 7. Bolstered resubstitution error estimates for LDA, CART, and KNN.

Here, the opposite is observed when a linear SVC is used with different values of penalty parameters (Table 3). Irrespective of the amount of data, the error estimates for each of the classifiers are around 50% which makes it unsuitable to be used as an ideal classifier for the separation of data between the classes as it showcases an ideal case of underfitting.

3.2. Bolstered leave one out (BLOO) error estimation

As the name suggests, in this type of error estimation, the classifiers are trained on each subset separately. Every subset contains all the data points belonging to one particular class except one which is used as the test datapoint. Therefore, in a binary classification problem, if m datapoints belong to class 0 and n datapoints belong to class 1, there will be a total of m-1 subsets for class 0 and n-1 subsets for class 1; and the classifiers will be trained separately on all of them. In the end, an aggregate of the error on each of these subsets is calculated to find out the total Bolstered Leave One Out Error estimate for each of these classifiers.

Table 3
Bolstered resubstitution error estimates for linear SVM with varying values of penalty parameter (C).

Sl no	N	C = 1	2	3	5	10	20	100
1	20	53.57	53.57	53.32	53.57	44.36	53.6	52.8
2	30	53.48	53.48	53.48	53.48	53.5	53.5	52.8
3	40	54.38	54.38	53.38	54.38	53.3	54.4	52.8
4	50	53.52	53.17	53.17	53.52	52.8	47.7	53
5	60	53.48	53.48	53.48	53.48	53.5	53.4	52.4
6	70	54	53.82	53.82	53.96	50.4	52.5	53.9
7	80	53.91	53.95	53.9	53.96	52.2	53	53.8
8	90	53.91	51.86	51.56	53.9	53.9	51.8	53.9
9	100	53.84	53.8	53.8	53.84	53.8	53.8	53.8

A general form of the Bolstered Leave One Out Error Estimator can be formulated as follows:

$$A^i_j = \left\{ x \, \epsilon \, R^p | g\left(S^i_{n-1}, x\right) = j \right\}, \quad for \; j = \{0, 1\} \tag{4}$$

The error estimates for the linear classifiers are calculated similarly as in Eq. (3) over the decision regions defined above. For the non-linear classifiers where the integrals cannot be computed exactly, a Monte-Carlo approach is used to find out the error estimates.

The BLOO error estimates for LDA, KNN, and CART have been calculated for each of the classifiers and have been shown in Fig. 8.

From the above figure, it can be observed that as the size of the dataset increases from 20% to 100%, the value of the BLOO error estimate declines to less than 1%. This makes the BLOO error estimate an unreasonable metric to be used both in conjunction with either linear or non-linear classifiers. In the case of Linear SVM, the BLOO error estimate could not be calculated. This is because the BLOO error estimate is calculated on a single class of data at a time and aggregates the error estimates in the end, but SVM works on the principle of separating hyperplane between the classes which makes it impossible to train on a single class of data. Therefore, the third type of error estimate, known as the Semi-Bolstered Error Estimation, is proposed which is discussed in detail in the next sub-section.

3.3. Semi-bolstered resubstitution error estimation

Sometimes, in the case of small datasets, there is a high chance of increasing the optimism of the error estimate due to overfitting classification rules. So, it is not a good idea to spread the incorrectly classified datapoints. Therefore, to reduce the bias of the estimator, σ_i is assigned to 0 for the incorrectly classified datapoints. This increases variance because there is less bolstering, which has proven to improve the bias-variance trade-off when trained with non-linear classifiers. Therefore, the semi-Bolstered resubstitution error estimation technique has been implemented with the linear and non-linear classifiers stated above, and the results have been discussed in the tables below.

From Fig. 9, it is evident that the semi-Bolstered resubstitution error values for LDA, CART, and KNN are almost negligible (approaching zero). Therefore, it can be stated that owing to the small size of the dataset, even reducing the bias of the estimator by assigning the Standard Deviation of the Bolstering kernel for the incorrectly classified data points to zero does not work as they tend to overfit the classifiers.

Contrary to the above, when the same dataset was trained using Linear SVC with Semi-Bolstered resubstitution as the error metric, the values of error estimates seemed reasonable (Table 4) It can be observed that when Linear SVM was trained using 100% of the data, the semi-Bolstered error estimates remained constant irrespective of the values of the penalty parameter used. Therefore, it was decided to proceed with Linear SVC with the penalty parameter set to 1 and Semi-Bolstered resubstitution as the error estimate.

Based on the value of this Machine Learning model which either outputs 0 or 1, a set of recommended rules have been prescribed to regulate the nutrient parameters in the aquaponic solution which have been stated in Table 5.

4. Discussion

The main motivation behind using this entire approach is designing a recommendation system so that the users of the hydroponic set-ups get an alert on the appropriate levels of nutrient concentrations that should be maintained to obtain the maximum yield. To achieve this, the training dataset has been used to design the approach. From the values of the pre-recorded observations, a set of rules were developed for both the classes.

For the observations to which the output class was 0, a certain set of inferences can be made. The period in which we are most likely to receive this output is generally during the Spring and Summer months (January to August). During these months when the temperatures increase (to about 80 F in a typical year), the concentration of nutrients in water also increase due to evaporation and often results in nutrient imbalance in the aquaponic solution. The electrical conductivity of the solution declines fast, which is addressed by addition of potassium to maintain conductivity at 350 umhos/cm. During these months, the growth of both the crop and the fish is much slower when compared to the winter months; therefore, caution needs to be exercised to balance the nutrient concentrations of the solution while trying to keep the hardness of the solution as low as possible. Another important consideration is the concentration of magnesium in the solution as it is one of the important constituents of fish culture. Since the goal during the spring and summer months is mostly to sustain the fish rather than grow them, a concentration of 1 ppm of magnesium is enough for the aquaponic solution. Lastly, the concentration of sodium should be kept at a maximum of 60 ppm since sodium increases water salinity, which directly impacts the growth of green leafy vegetables.

Similarly, for the observations with output class 1, irrigation is likely done during autumn and winter months (September to December). During these months as the atmospheric temperatures fall to an average of 50 F, evaporation slows down and as a result, the conductivity and

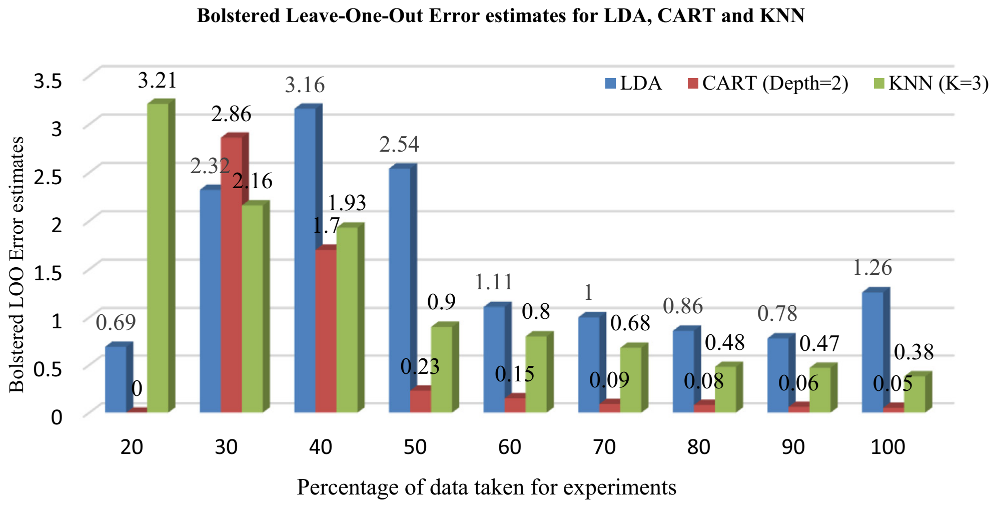


Fig. 8. Bolstered leave-one-out error estimates for LDA, CART, and KNN.

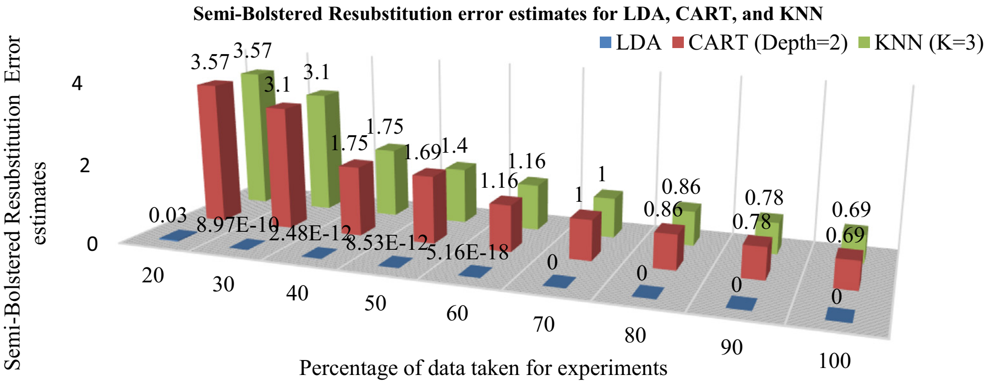


Fig. 9. Semi-bolstered resubstitution error estimates for LDA, CART, and KNN.

hardness of the solution can be maintained at 1970 umhos/cm and 140 ppm of calcium carbonate, respectively, which is important given the higher yield levels (about double) obtained during these months compared to the rest of the growing season. Any increase in these properties can increase the hardness of the solution, resulting in eutrophication (Glibert et al., 2005). To control this and maintain the hardness of the solution within 140 ppm, anoxic mixers need to be installed to denitrify the wastewater with the use of bacteria, which breaks down the nitrates to release oxygen and nitrogen gas (Peng et al., 2008).

Based on this set of recommendations, two different types of temperature-controlled environments were set up, one with a temperature of 75 F to replicate Texas summers and the other with a temperature of 45 F to replicate Texas winters. The nutrient concentrations of the aquaponic solution were regulated as per the above recommendations and the observations are listed as follows. In the first tank, lettuce was grown using the recommended nutrient concentrations. It was observed that the harvest time for lettuce decreased to 4 weeks as compared to 6 to 8 weeks when cultivated using traditional aquaponic techniques. There was also a significant increase in the yield and size of lettuce and some were as large as 40–45 in. in diameter which were significantly bigger compared to the ones grown in the soil or the unregulated aquaponic solution. Kale was grown in the second tank and similar to that of lettuce, the growth period was reduced from 12 weeks to 4.5 weeks.

Kledal et al. have stated that the pH needs to be maintained between 6.5 and 8 by maintaining a balance between the chemical concentrations of nitrogen, phosphorus, potassium, sulfur, calcium, magnesium and heavy metals namely iron, manganese, boron, copper, zinc and molybdenum to facilitate growth of microbes in the aquaponic solution which is vital for nutrient uptake by plants (Kledal et al., 2019). However, the ideal range of nutrients that should be maintained in the solution has not been discussed by Kledal et al., and our work can be viewed as a quantification to facilitate the findings that were made in this paper.

Table 4
Semi-bolstered resubstitution error estimates for linear SVM with varying values of penalty parameter (C).

Sl no	N	C = 1	2	3	5	10	20	100
1	20	15.2	17.94	22.28	22.37	17.7	13.8	18.2
2	30	16.3	18.66	16.34	11.76	21.3	18.9	14.1
3	40	14.04	21.06	17.57	21.97	14.3	21.1	24.7
4	50	19.71	22.54	12.69	18.39	19.8	15.6	23.1
5	60	17.44	12.85	19.87	17.47	23.3	12.7	17.4
6	70	19	16.01	21	18.06	17	16.1	20
7	80	19.13	19.13	17.39	19.13	19.1	20	20
8	90	19.53	20.31	18.7	17.97	17.2	19.5	18.8
9	100	18.18	18.18	18.18	18.18	18.2	18.2	18.2

Likewise, Nozzi et al. have compared the nutritional quality of lettuce, mint and mushroom herbs grown in three different aquaponic set-ups with varying amounts of additional macronutrients and nutrient uptake from fish feed, but there has been no concrete explanation for how nutrient addition in each case supplemented the growth of these plants (Nozzi et al., 2018). In our case, a proper description of how the proposed recommendations of the Machine Learning model affected the quality of the crops in the aquaponic solution, as compared to the baseline model, has been elaborately discussed.

This is the first study of its kind to use AI for optimizing nutrient concentrations in aquaponic solution. Our proposed method is much superior to the conventional methods of growing plants in aquaponics as it emphasizes on quantifying the regulation of nutrients and routine monitoring according to the month in which the crops are grown. The number of nutrients that need to be regulated in the aquaponic solution as mentioned before were chosen using feature selection and dimensionality reduction techniques, and we finalized five most important chemical parameters from the initial set of 20 parameters, which resulted in reducing the cost of nutrients by around 75%. Further, the AI model we developed here is scalable and could be implemented in aquaponic operations at various sizes. Our model also overestimates the possibility of toxicity to the plants and fish, which in fact ensures produce quality. The most important issue which has been addressed by automation in aquaponics is the cost incurred due to human labor as aquaponics only involves an initial cost of setting up the farm and water purification units for each of the tanks, with the only recurring cost being the cost incurred with utilities.

However, the approach presented here does have some limitations. The sparse data makes it difficult to gauge the structure of the data for applying any Bayesian statistical approaches. This can be compensated with an increase in the volume of the dataset, which needs to be addressed in future. Another limitation is the weather conditions in which the observations were recorded, i.e. all the three aquaponic farms were located in Southeast Texas, with warm summers and mild winters. Addition of data from geographical locations with much more variable weather conditions is expected to further improve the robustness of the model presented here.

Table 5
Recommendation system for each predicted class.

Nutrient concentrations	Class 0	Class 1
Magnesium (ppm)	1	12
Sodium (ppm)	60	320
Conductivity (umhos/cm)	350	1970
Hardness (ppm of calcium carbonate)	27	140

5. Conclusion

From the above experimental results, it can be concluded that the Semi-Bolstered resubstitution Error estimation technique works best with Linear SVC as the classifier with the value of penalty parameter set to 1. A set of recommendations have also been prescribed for each of the predicted classes, which have been discussed in detail in this paper. It has been verified experimentally as to how the proposed recommendations positively impacted plant and fish growth in a single system. In the future, more attention would be paid on how the concentration of heavy metals affects the growth of plants in aquaponic environments. Moreover, the success of the classification technique presented here can be extended to other data-depleted domains, which is another contribution of this paper.

Credit authorship contribution statement

Sambandh Bhusan Dhal: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Muthukumar Bagavathiannan:** Conceptualization, Validation, Writing – review & editing. **Ulisses Braga-Neto:** Methodology, Software, Investigation, Project administration. **Stavros Kalafatis:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I am grateful to Ms. Sharon Wells, owner of Aquatic Greens Farm, Bryan; Mr. Robert Wolff, owner of Wolff Family Farms, Caldwell; and Mr. Joe Leveridge, owner of Texas US Farms, Grimes County for their cooperation and providing access to their hydroponic farms for experimentation. Their insights have been equally helpful in designing the approach that has been implemented here.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscipl. Rev. Comp. Stat. 2 (4), 433–459. https://doi.org/10.1002/wics.101.

Ahmed, U., Lin, J.C.-W., Srivastava, G., Djenouri, Y., 2021. A nutrient recommendation system for soil fertilization based on evolutionary computation. Comput. Electron. Agric. 189, 106407. https://doi.org/10.1016/j.compag.2021.106407.

Arvind, C.S., Jyothi, R., Kaushal, K., Girish, G., Saurav, R., Chetankumar, G., 2020. Edge computing based smart aquaponics monitoring system using deep learning in IOT environment. 2020 IEEE Symposium Series on Computational Intelligence (SSCI) https://doi.org/10.1109/ssci47803.2020.9308395.

Braga-Neto, U., Dougherty, E., 2004. Bolstered error estimation. Pattern Recogn. 37 (6), 1267–1281. https://doi.org/10.1016/j.patcog.2003.08.017.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40 (1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Concepcion II, R.S., Alejandrino, J.D., Lauguico, S.C., Tobias, R.R., Sybingco, E., Dadios, E.P., Bandala, A.A., 2020. Lettuce growth stage identification based on phytomorphological variations using coupled color superpixels and multifold watershed transformation. Int. J. Adv. Intellig. Inform. 6 (3), 261. https://doi.org/10.26555/ijain.v6i3.435.

Cox, M.A., Cox, T.F., 2008. Multidimensional scaling. Handb. Data Visualiz., 315–347 https://doi.org/10.1007/978-3-540-33037-0_14.

Dhal, S.B., Jungbluth, K., Lin, R., Sabahi, S.P., Bagavathiannan, M., Braga-Neto, U., Kalafatis, S., 2022. A machine-learning-based IoT system for optimizing nutrient supply in commercial aquaponic operations. Sensors 22, 3510. https://doi.org/10.3390/s22093510.

Dietterich, Tom, 1995. Overfitting and undercomputing in machine learning. ACM Computing Surveys 27 (3), 326–327. https://doi.org/10.1145/212094.212114.

Glibert, P., Seitzinger, S., Heil, C., Burkholder, J.A., Parrow, M., Codispoti, L., Kelly, V., 2005. The role of eutrophication in the global proliferation of harmful algal blooms. Oceanography 18 (2), 198–209. https://doi.org/10.5670/oceanog.2005.54.

Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom. Intell. Lab. Syst. 83 (2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007.

Hesterberg, T., 2011. Bootstrap. Wiley Interdiscipl. Rev. Comp. Stat. 3 (6), 497–526. https://doi.org/10.1002/wics.182.

Jensen, M.H., 1997. Hydroponics. HortScience 32 (6), 1018–1021.

Jones Jr., J.B., 2016. 2016 Hydroponics: A Practical Guide for the Soilless Grower. CRC Press.

Karimanzira, D., Na, C., Hong, M., Wei, Y., 2021. Intelligent Information Management in aquaponics to increase mutual benefits. Intell. Inf. Manag. 13 (01), 50–69. https://doi.org/10.4236/iim.2021.131003.

Kledal, P.R., König, B., Matulić, D., 2019. Aquaponics: the ugly duckling in organic regulation. Aquapon. Food Prod. Syst. 487–500. https://doi.org/10.1007/978-3-030-15943-6_19.

Köppen, M., 2000. The curse of dimensionality. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5). 1, pp. 4–8.

Lauguico, S.C., Concepcion II, R.S., Alejandrino, J.D., Tobias, R.R., Macasaet, D.D., Dadios, E.P., 2020. A comparative analysis of machine learning algorithms modeled from machine vision-based lettuce growth stage classification in Smart Aquaponics. Int. J. Environ. Sci. Dev. 11 (9), 442–449. https://doi.org/10.18178/ijesd.2020.11.9.1288.

Likas, A., Vlassis, N., Verbeek, J.T., 2003. The global K-means clustering algorithm. Pattern Recogn. 36 (2), 451–461. https://doi.org/10.1016/s0031-3203(02)00060-2.

Lobanov, V.P., Combot, D., Pelissier, P., Labbé, L., Joyce, A., 2021. Improving plant health through nutrient remineralization in aquaponic systems. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.683690.

Mahanta, S., Habib, M.R., Moore, J.M., 2022. Effect of high-voltage atmospheric cold plasma treatment on germination and heavy metal uptake by soybeans (glycine max). Int. J. Mol. Sci. 23, 1611. https://doi.org/10.3390/ijms23031611.

Maleki-Kakelar, M., Azarhoosh, M.J., Golmohammadi Senji, S., Aghaeinejad-Meybodi, A., 2021. Urease production using corn steep liquor as a low-cost nutrient source by Sporosarcina pasteurii: Biocementation and process optimization via Artificial Intelligence Approaches. Environ. Sci. Pollut. Res. 29 (10), 13767–13781. https://doi.org/10.1007/s11356-021-16568-6.

Nozzi, V., Graber, A., Schmautz, Z., Mathis, A., Junge, R., 2018. Nutrient management in aquaponics: comparison of three approaches for cultivating lettuce. Mint Mushroom Herb. Agronomy. 8 (3), 27. https://doi.org/10.3390/agronomy803002.

Peng, Y., Hou, H., Wang, S., Cui, Y., Zhiguo, Y., 2008. Nitrogen and phosphorus removal in pilot-scale anaerobic-anoxic oxidation ditch system. J. Environ. Sci. 20 (4), 398–403. https://doi.org/10.1016/s1001-0742(08)62070-7.

Pillay, T.V.R., 2003. Aquaculture and the Environment. 2003. John Wiley & Sons.

Pillay, T.V.R., Kutty, M.N., 2009. Aquaculture: Principles and Practices. No. 2. Blackwell Publishing, p. 2009.

Ponce, H., Cevallos, C., Espinosa, R., Gutierrez, S., 2021. Estimation of low nutrients in tomato crops through the analysis of leaf images using machine learning. Spec. Issue: Blockchain Artific. Intellig. Appl. 1 (2). https://doi.org/10.37965/jait.2021.0006.

Provin, T.L., Pitt, J.L., 2002. Description of Water Analysis Parameters.

Rau, J., Sankar, J., Mohan, A.R., Das Krishna, D., Mathew, J., 2017. IoT based smart irrigation system and nutrient detection with 618 disease analysis. 2017 IEEE Region 10 Symposium (TENSYMP), 2017, pp. 1–4 https://doi.org/10.1109/TENCONSpring.2017.8070100.

Roberto, K., 2003. How-to hydroponics, Futuregarden. Inc. 2003.

Shafique, R., Mehmood, A., Ullah, S., Choi, G.S., 2019. Cardiovascular disease prediction system using extra trees classifier. https://doi.org/10.21203/rs.2.14454/v1.

Shao, Y., Lin, J.C.-W., Srivastava, G., Guo, D., Zhang, H., Yi, H., Jolfaei, A., 2021. Multi-objective neural evolutionary algorithm for combinatorial optimization problems. IEEE Trans. Neural Networks Learn. Syst. 1–11. https://doi.org/10.1109/tnnls.2021.3105937.

Sima, C., Braga-Neto, U., Dougherty, E.R., 2004. Superior feature-set ranking for small samples using bolstered error estimation. Bioinformatics 21 (7), 1046–1054. https://doi.org/10.1093/bioinformatics/bti081.

Timsina, J., Dutta, S., Devkota, K.P., Chakraborty, S., Neupane, R.K., Bishta, S., Amgain, L.P., Singh, V.K., Islam, S., Majumdar, K., 2021. Improved nutrient management in cereals using nutrient expert and machine learning tools: productivity, profitability and nutrient use efficiency. Agric. Syst. 192, 103181. https://doi.org/10.1016/j.agsy.2021.103181.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2 (1-3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

Yadav, A., Thakur, U., Saxena, R., Pal, V., Bhateja, V., Lin, J.C.-W., 2022. AFD-Net: apple foliar disease multi classification using deep learning on plant pathology dataset. https://doi.org/10.21203/rs.3.rs-1158879/v1.