

A Machine Learning Framework for Optimizing Biofloc-Based Aquaculture: A Case Study from Bangladesh

Sraboni Ghosh Joya^a and Nila Sultana^b

^{a,b} Department of Computer Science and Engineering (CSE), City University
Corresponding Author: Sraboni Ghosh Joya

Abstract

Aquaculture is in great demand for developing sustainable and highly efficient aquaculture system in the world, especially in the countries like Bangladesh where fish is the major source of protein for almost 79% people of the country. To tackle these aquaculture challenges, a new technology has surfaced in the last few years that has become known as Biofloc technology. A novel technology has developed over the last few years, now termed Biofloc technology that may offer a more sustainable and efficient way of reducing environmental impacts, increasing resource utilization efficiency and increasing production yields. In particular, a new machine learning technique called Impact Learning is introduced and tested against well-established machine learning techniques, such as Support Vector Machines, Decision Trees, Naïve Bayes, Random Forest, and Multilayer Perceptron. The preprocessed datasets were used to train the models, and the test datasets were used to evaluate the models using various metrics, including accuracy, precision, recall, F1 score, cross-validation score, etc. Random Forests and Decision Trees were the best models overall. The outcomes showcase the capabilities of data-driven approaches to improve the precision of predictions and offer operational insights, which support sustainable aquaculture practices, including the application of Impact Learning.

Keywords: Biofloc Technology, Aquaculture, Machine Learning, Impact Learning, Sustainable Fisheries.

Date of Submission: 27-09-2026

Date of Acceptance: 05-10-2026

I. Introduction

Aquaculture is now one of the largest and fastest-growing food production industries in the world and plays an important role in meeting the world's demand for high-quality protein. The amount of fish production from aquaculture now exceeds 50% of the fish consumed globally, and will continue to increase with decreasing fish stocks of the wild population and growing world population [1]. The fisheries sector in the country is playing a key role in food supply and employment, and about 80% of animal protein is consumed in the form of fish in Bangladesh [2].

Although the technology has great potential, the conventional aquaculture in Bangladesh is confronted with a few major issues. These include overwatering, accumulating wastes, high feed conversion ratio, outbreaks of disease, and environmental damage resulting from effluent discharge [3]. To overcome these challenges, Biofloc Technology (BFT) has been developed as an alternative that is sustainable and resource-efficient. BFT is based on the principle that the organic wastes of fish, primarily feed waste and excrement, can be utilized by microbes to produce microbial biomass that can be fed to fish as an additional nutrient source. This helps to minimize environmental impact and also enhances feed efficiency and water quality [4,5].

From the perspective of low investment and low technology, low environmental pollution, and high profitability, BFT has attracted a lot of interests for adoption. BFT is gaining popularity as it is a low tech farming method with low technology, low environmental pollution, and high profitability, particularly in the developing countries where the infrastructure and resources are limited and challenging for farmers to do farming. It can produce fish in high density with low water exchange rate in a small space and is appropriate for the water-scarce and high demand areas like Bangladesh [6]. However, the process and maintenance of biofloc systems are still difficult to operate and maintain because of its multiple dynamic parameters like temperature, pH, dissolved oxygen, ammonia and microbial activity. The complexity requires powerful tools to monitor and optimise, in real time and based on data.

In recent years, machine learning (ML) has emerged as a revolutionary technology for many aquaculture purposes, such as water quality prediction [7], fish activity monitoring [8] or disease detection [9]. They can learn complex patterns when applied to high dimensional data and make accurate prediction in real time which may be useful in precision aquaculture system. Nevertheless, there is limited study on biofloc aquaculture system with the addition of machine learning models especially in the developing countries.

In this study, a complete framework was developed to predict and optimize the biofloc performance by machine learning models with empirical data from various aquaculture farms in Bangladesh to overcome the aforementioned issues. In particular, it introduces a novel model, called Impact learning, a powerful supervised learning approach that takes into account the impact of competitive features on the decision-making process [10]. The results are contrasted with the classifiers SVM, DT, NB, RF and MLP.

Considering this, the main objective of this study is to systematically evaluate these models according to various performance measures such as accuracy, F1 score, precision, recall, and cross-validation to conclude the most effective algorithm to improve productivity, cost-effectiveness, and sustainability in aquaculture of biofloc. In conclusion, the results are of great value to researchers and stakeholders in the field of data-driven aquaculture, urging for sustainable fish farming practices in Bangladesh and other parts of the world.

II. Literature Review

The aquaculture sector is evolving rapidly, as the need for sustainable protein sources is growing, and smart farming technologies are being developed in conjunction. In particular, Biofloc Technology (BFT) and machine learning technologies (with the support of predictive systems) have offered new prospects to improve resource efficiency and aquaculture output. In this section, a review of prior studies was performed on three primary topics: Development and performance of Biofloc system, technological solutions for aquaculture difficulties and machine learning applications for water quality prediction and yield estimation.

Biofloc Technology in Aquaculture

Biofloc Technology (BFT) has been proven as an innovative approach in aquaculture that has been developed to provide solutions for intensive fish farming in limited water resources. The principle of BFT was first introduced by Avnimelech [11], whereby organic waste is converted to high-protein biomass by microbes, which improves the water quality as well as supplements the fish feed. This system significantly decreases the amount of water that needs to be exchanged, and the feed conversion ratios are also significantly enhanced. The technology has been implemented in diverse settings, and research is emerging on various species that have shown the advantages of using the technology, including tilapia and *Penaeus vannamei* [12].

However, the implementation of BFT needs to be monitored and tuned continuously for some important water quality parameters like ammonia, pH, dissolved oxygen, and carbon-to-nitrogen ratio (C/N). Betanzo-Torres et al. [13] found that poor parameter control and absence of technical know-how are significant challenges for wider adoption in developing nations. In addition, Rashid et al. [14] stressed the necessity of smart water quality monitoring, so that BFT can be scaled up in countries such as Bangladesh.

Technological Advancements in Aquaculture

Today, the Internet of Things (IoT) and the combination of sensor networks and data analysis have created a new trend: Precision Aquaculture, intelligent decision-making based on real-time data. Føre et al. [16] presented an overall system design for Smart Aquaculture Systems composed of sensors that monitor the environment and actuators that control feed and aeration. Real-time sensing of critical variables is successful with IoT-based platforms. In their work, Huan et al. [17] proposed a water quality monitoring system using NB-IoT, in which they demonstrated better responsiveness and remote monitoring. Likewise, Amrita et al. [18] implemented an autonomic water quality management system with integrated water quality sensors and notification nodes for alerting to faults. But most of these technologies offer descriptive information. The predictive analytics layer, which could be used for anticipatory responses (such as modifying the C/N ratio before ammonia peaks), has not been extensively studied in BFT. This is helping the development of machine-learning models capable of working with multi-dimensional data and predicting conditions in aquaculture.

Machine Learning in Aquaculture Prediction

Machine learning (ML) has been found useful in aquaculture for predicting fish health, optimizing feed, and measuring environmental conditions.

The Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Artificial Neural Networks (ANNs) are some of the algorithms used with success that can identify non-linear relationships and hidden patterns in complex datasets [19].

In Sunayana et al. [20], Artificial Neural Networks were used to predict the hardness of groundwater based on multiple learning algorithms by incorporating hydro-geochemical variables. These methods have been adopted in other applications, such as dissolved oxygen prediction [21] and water quality classification [22]. The studies

show that ML can be adapted to environmental modelling, especially in the presence of noise and/or many variables, as is the case when working with aquaculture data.

In recent years, Mustapha et al. [23] pointed out that the combined application of cloud computing and IoT with AI could be a breakthrough in sustainable aquaculture, which they called the CIA framework (Cloud, IoT, AI). The integration enables large-scale data-driven aquaculture systems. However, most of them are based on traditional ML methods and do not consider the feature interaction in the domain. To overcome the above shortcomings, Kowsher et al. [24] introduced a supervised learning approach called Impact Learning that was used to account for the competitive effects of the input attributes, particularly in dynamic environments where the effects of several variables overlap. This is a relatively new approach, but it has not been used in aquaculture settings, which would be a great advance in the methods used for BFT models.

Table 1: Summary of Related Work and Contributions of the Present Study

Study	Focus Area	Techniques/Models Used	Limitations/Scope	Relevance to Current Work
Avnimelech (2015) [11]	Introduction of Biofloc Technology	Microbial conversion of waste to protein	Conceptual, lacks computational integration	The foundation of BFT is applied in this study
Rashid et al. (2021) [14]	IoT for Biofloc water quality	Sensor-based monitoring	No predictive modeling or ML usage	Highlights the need for predictive analytics in BFT
Betanzo-Torres et al. (2020) [13]	BFT adoption in developing regions	Surveys and qualitative analysis	No computational or algorithmic insights	Supports context-specific challenges in BFT deployment
Føre et al. (2018) [16]	Precision fish farming framework	Sensors, automated systems	No machine learning for optimization	Motivates ML-based operational enhancement
Huan et al. (2020) [17]	Water quality monitoring	NB-IoT-based system	Focused only on monitoring, not prediction	Basis for using real-time data in predictive systems
Sunayana et al. (2020) [20]	Groundwater hardness prediction	ANN, ensemble models	Not related to aquaculture, but a strong ML benchmark	Demonstrates ML potential in environmental modeling
Mustapha et al. (2021) [23]	Sustainable aquaculture vision	Cloud + IoT + AI (CIA)	Theoretical integration without algorithm testing	Validates integration of AI in aquaculture systems
Kowsher et al. (2020) [24]	Introduction of Impact Learning	Impact-Learning algorithm	No application to aquaculture	This study introduces it into BFT prediction

Summary and Research Gap

Although previous research has demonstrated the potential of machine learning for environmental prediction and the advantages of biofloc systems, there is scarce research on machine learning integrated with biofloc systems, particularly with local datasets from countries such as Bangladesh. Furthermore, the lack of literature on the use of Impact Learning in aquaculture or comparing it to other common models. This work is a contribution to this area because:

- Adopting the new ICT concept, and
- Comparing it with 5 conventional models (SVM, DT, NB, RF, MLP),
- We examined the relationship between plantains and their root yields, using actual data from various farms in Bangladesh.
- Assessing performance using measurement and testing procedures that stress accurate measurement.

This review of literature shows that significant advancement has been made, and biofloc aquaculture systems are being developed and implemented, and digital technologies are being increasingly used in aquaculture management. Biofloc technology is known to improve the sustainability, reduce the operational cost, and enhance the water quality by converting the waste materials to usable biomass [11, 14]. Furthermore, the advent of monitoring systems using Internet of Things (IoT) principles has provided the basis for real-time monitoring in the aquaculture context [17, 18], which allows for improved decision-making and environmental control. Most of these systems, however, are descriptive, and make observations but do not make predictions.

In aquaculture, machine learning algorithms have several potential uses, including predicting water quality parameters, monitoring fish health, and optimizing feeding and watering practices [19, 20, 22]. These investigations conclude the relevance of supervised learning algorithms like Random Forest, Support Vector Machines or Artificial Neural Networks in the identification of non-linear relationships in this complex matrix of environmental information. Despite this, most of the applications currently available are restricted to specific environmental variables or, in particular, to generic datasets, and are oriented toward a minimum number of domain-specific issues like those of biofloc systems.

Furthermore, despite successful efforts in the technical aspects of smart aquaculture, few studies have explored regional differences and practical constraints in resource-poor contexts such as Bangladesh. Numerous

models developed in previous research rely on global data or laboratory settings, restricting their applicability and quality in translating into biofloc farming systems that are not only varied in their water chemistry, microbial metabolism, and farming practices, but are also widely used by farmers in the field.

A key unaddressed research challenge is the lack of integration of novel or complex machine learning methods that are specific to a competitive feature environment. There is no current research on the aspects of aquaculture or environmental modelling that the recently proposed Impact Learning model proposes to consider, that is, the interaction and competition between the features during training. Its utility in BFT may provide a more powerful and interpretable prediction, particularly in dynamic systems where the water quality measurements are associated in complex ways.

This study aims to fill these gaps by:

- The first time the introduction and test of the Impact Learning algorithm in a biofloc aquaculture context.
- Comparing the performance of the five well-established supervised machine learning model namely SVM, DT, NB, RF, and MLP, to predict the impact of learning and testing their performance on a locally collected data set from 50 farms in Bangladesh.
- Full range of water quality and operating parameters reflective of the actual BFT systems.
- Testing all models with all statistical performance measures (accuracy, precision, recall, F1-score, cross-validation) in a statistically reliable manner for generalization.
- Incorporating a scalable, data-based framework that can be adopted by aquaculture producers and policy makers for a more intelligent and sustainable fisheries production in the developing world.

III. Methodology

This section outlines the end-to-end experimental framework from data acquisition to data preprocessing, model development, model evaluation, and validation procedures. The aim is to design and test supervised machine learning models for recommending the best operating practices in biofloc systems based on the empirical data from various farms in Bangladesh. It is analyzed with classifiers such as support vector machines, decision trees, random forest, naive bayes, and multi-layer perceptron, and impact learning. In this section, these algorithms are also explained. These algorithms are employed to forecast the best, optimal, and most economically viable machine learning models in the biofloc, resulting in better returns. This model is then tested with the proper evaluation metric. In addition, different performance metrics are used in data evaluation, including confusion metrics, correlation matrix, accuracy, precision, and recall, all of which are always supportive of our study.

Workflow Diagram

The methodology consists of the following steps:

- Data Collection in 50 aquaculture farms of Bangladesh.
- Data Preprocessing (Cleaning, Normalizing, Feature Engineering).
- Six machine learning algorithms were used for Model Development, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Multilayer Perceptron (MLP), and the proposed Impact Learning (IL) algorithm.
- Evaluation of performance using cross-validation and metrics like accuracy, precision, recall, and F1 Score.
- Model Comparison, Choosing the best and most generalizable classifier.

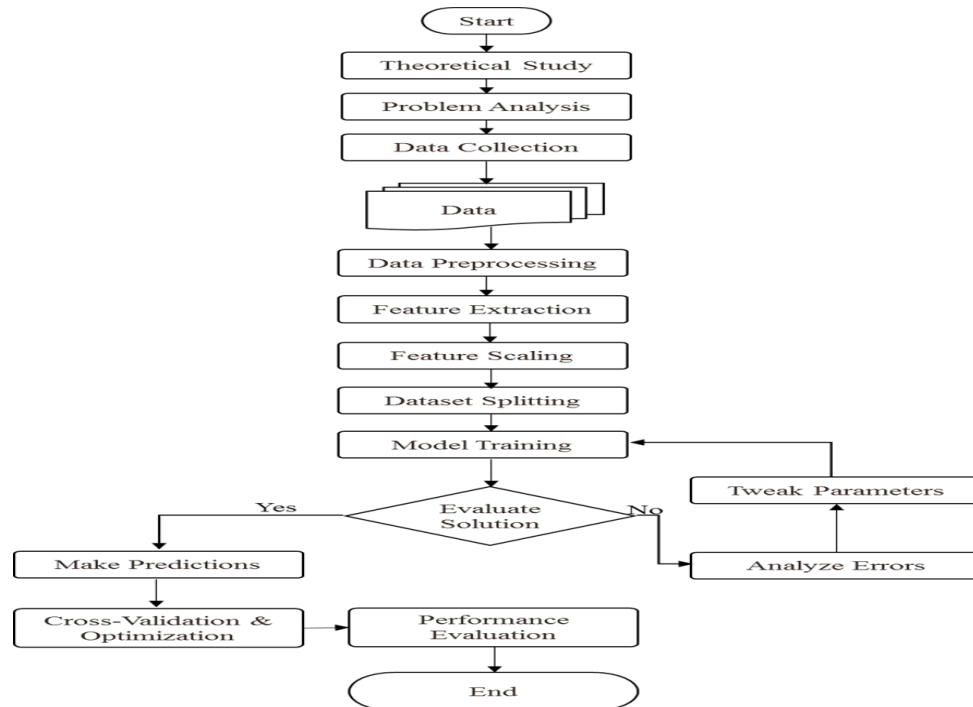


Figure 1. Workflow diagram.

This diagram is a simplified depiction of the entire system. It gives a clear idea of the process of forecasting. Data was collected from various biofloc farms in the country first. Once the data was collected, the unknown, duplicate, and missing data were removed. Thus, the following step would be data preprocessing. The data is analyzed with the help of data mining tools. We also scaled the function so that the model can be trained and converge faster. Standardized methods of feature scaling are used. Then, we begin designing our model. We have developed a supervised machine learning model that is trained with the already preprocessed data, and the trained data we use to predict, which is the best, ideal, and most profitable machine learning model in biofloc. In this experiment, different machine learning techniques are employed, such as impact learning. The last stage is to determine the accuracy of these models with different performance metrics. The best accuracies of these models were selected by comparison.

Data Collection and Description

Field surveys were conducted at 50 aquaculture farms across 11 districts in Bangladesh: Savar, Dhaka, Munshiganj, Dinajpur, Gaibandha, Gazipur, Narayanganj, Cumilla, Mymensingh, Narsingdi, and Sirajganj. A total of 200 floc samples were collected, focusing on key cultivars including Tilapia (*Oreochromis niloticus*), Singhi (*Heteropneustes fossilis*), and Pabda (*Ompok bimaculatus*). Table 2 lists the selected physicochemical and operational variables.

Table 2: Dataset Description.

Variable Name	Type	Unit
Temperature	Numeric	°C
Dissolved Oxygen	Numeric	mg/L
pH	Numeric	scale (0–14)
Salinity	Numeric	ppt or %
Water Color	Categorical	-
Probiotics	Numeric	CFU (colony-forming units)
TDS	Numeric	mg/L

Variable Name	Type	Unit
Ammonia (NH ₃)	Numeric	mg/L
Nitrate (NO ₃ ⁻)	Numeric	mg/L
Nitrite (NO ₂ ⁻)	Numeric	ppm

Data Preprocessing

One of the most critical states of the implementation module. We've collected a great deal of data from various sources, and some of that data is garbage, and some of that is null. The dataset may contain multiple (ambiguous), duplicate, or missing values for most values. It may also be confusing because some data may have big dimensions, and it may require a long time to train. Hence, data pre-processing is an important step in this process. The pre-processing has been done in Python (pandas, NumPy, scikit-learn). It involved:

- Treatment of missing data: Imputation of missing data by the k-Nearest Neighbors (k=5) for continuous variables and by the mode for categorical variables.
- Duplicate removal – All the duplicate entries were removed according to composite key matching.
- Outlier Detection: Applied IQR-based filtering and z-score normalization.

Feature Scaling

The feature scaling is one of the most crucial changes that needs to be carried out on the data. The basic rule of machine learning is that the algorithms are not even good at identifying differences within different numerical range scales. For this reason, there is a need to scale the data. Standardization is used to finish this. The standardization process is fairly straightforward. Subtract the mean from the values for each attribute, the first step. Next, divide by the variance to obtain a distribution of unit variance (unit standardized data will always have a mean of zero). This is the most effective approach since it's more resistant to the impact of outliers.

$$X_{\text{stand}} = \frac{X - X'}{\sigma} \quad (1)$$

Splitting Dataset

It is not advisable to access all info at once. The data set is typically split into two sets. However, the data sets must be divided into training and testing sets prior to model analysis for data mining. The information in a data set is frequently divided into two sections: Training is done in the first section, while testing is done in the second. To assist guarantee that the training and test sets are comparable, Analysis Services will employ a random sample of the data. The effects of data inconsistencies can be reduced and a deeper understanding of the model's characteristics can be achieved by using comparable data for training and testing. The training set makes use of most of the data. Seventy percent of the data was used to train the models. Thirty percent still need to be preserved for testing.

Model Development

This is the most important one, the development of the models that will be used to predict the phenomenon. 6 machine learning algorithms have been used for prediction. Here are the algorithms like Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact learning. In the following, all of these algorithms are discussed:

Support Vector Machine

The Supervised Learning Model, Supervised Learning Algorithms, are used for data analysis for Regression Analysis and Classification. The Supervised Learning Model is called the Supervised Vector Machine. It constructs the model as it learns from a set of data that has been previously divided into 2 sets for training. The category of a new data point has to be classified by an SVM algorithm. As a result, SVM can be considered a non-binary linear classifier. In addition to classifying items, an SVM method should maximize the space on a graph between each object. With SVM, the data can be separated into categories even if it cannot be separated in the original feature space. SVM maps the data into a high-dimensional feature space that allows it to categorize data points, even if they are not linearly separable in that original feature space. The data are changed to allow for the hyperplane representation of the separator once a dividing line between the categories has been found. Thus, based on properties of new data, the group it should be assigned to can be estimated. For example, look at the following graph, which shows that there are two sets of data points.

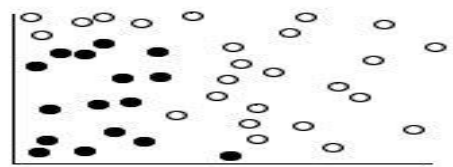


Figure 2. Scenario of the original dataset.

After the transformation, the boundary between the two categories can be established using a hyperplane, as seen in the image below.

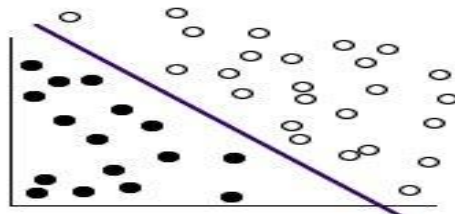


Figure 3. Scenario of transformed data.

The kernel function refers to the mathematical operation employed for the transformation.

Decision Tree

Decision Trees are versatile Machine Learning algorithms that can perform many output tasks, including regression and classification. An internal node is a test on an attribute, a branch represents a result of the test, and a leaf node (terminal node) corresponds to a class label. It has a tree structure similar to a flowchart. This is shown in Figure 2.6. The complete procedure is outlined in the following algorithm:

Step 1: Choose the starting point S as the root node, as it contains all the data.

Step 2: Determine the most important attribute in the data-set using the Attribute Selection Measure (ASM).

Step 3: Partition the S into mutually disjoint subsets containing the best values for each of the attributes.

Step 4: Select and construct the decision tree node having the best attribute.

Step 5: Create additional decision trees recursively, starting with subsets of the data that were obtained in step 3. The last node is referred to as a leaf node after continuing this process until the nodes can no longer be categorized.

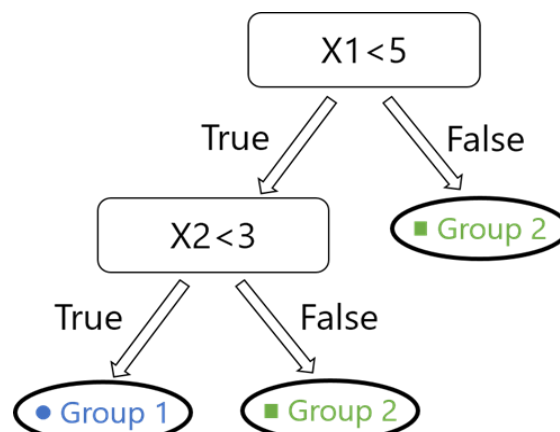


Figure 4. Example Of a Decision Tree.

Naïve Bayes

The Naive Bayes methodology is a simple approach to creating classifiers, or models that classify instances of issues described as feature vectors with a class label taken from a small set of possible values for the label. Not only is there not one way to train these classifiers, but there are multiple algorithms all based on the same principle: All naive Bayes classifiers assume that the value of one feature is independent from the value of any other feature, given the class variable. For instance, a round, red fruit about 10cm across is an apple. Naive Bayes, in its simplest form, is a conditional probability model. It takes as input a problem instance that needs to

be classified and is represented by a vector with certain features, $x = (x_1, \dots, x_n)$ (independent variables), and it assigns to this instance probabilities.

$$(C_k | x_1, \dots, x_n)$$

For each of k possible outcomes or classes C_k . Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k | x) = \frac{p(ck)p(x|ck)}{p(x)} \quad (4)$$

In plain English, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad (6)$$

The naïve Bayes classifier makes assumptions about feature distributions in its "event model." Multinomial and Bernoulli distributions are frequently used for discrete features, such as those that are present in classification tasks.

Random Forest

Random forest is a type of supervised learning algorithm. It puts together decision trees, often trained via "bagging," into a "forest." The basic idea of the bagging method is that the result is enhanced by aggregating multiple learning models. To find the greatest feature from a random subset of features, the Random Forest algorithm adds additional randomization when splitting a node.

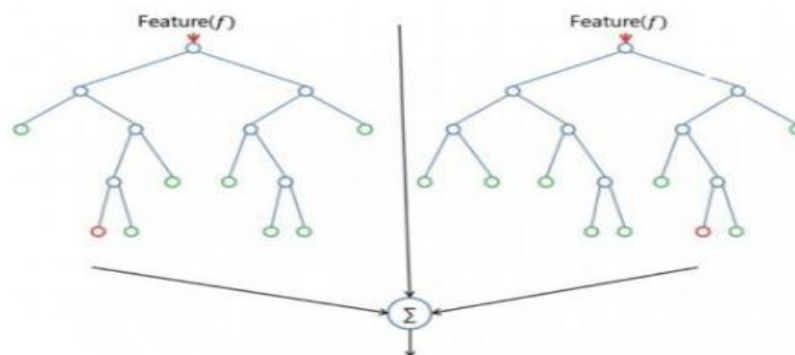


Figure 5. View of a random forest model.

The following steps can be used to illustrate the working process:

Step 1: Take K points at random from the practice set.

Step 2: Construct decision trees (Subsets) for the data points selected.

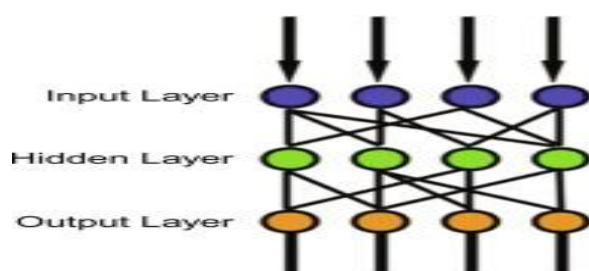
Step 3: Select the N -th node to be the decision tree node to build.

Step 4: Repeat Steps 1 and 2.

Step 5: Calculate the predictions for each decision tree on a set of new (unseen) data points, and put each new data point into the category that got the most votes.

Multilayer perceptron

A feedforward artificial neural network, called a Multilayer perceptron (MLP), generates a set of output vectors from a set of input vectors. The directed graph between the input and output layers of an MLP has multiple layers



of input nodes. Back propagation is used to train the network with MLP. MLP is a deep learning technique. MLP is widely used in research areas such as parallel distributed computing and computational neuroscience problems with supervised learning. Since MLPs can be made to approximate any continuous function, they can solve problems that cannot be solved by linear separation. The major application cases of MLP are Pattern Classification, Recognition, Prediction, and Approximation. The primary applications of MLP are

pattern categorization, pattern recognition, pattern prediction, and approximation. An input function u is applied to the input features, which calculates the weighted sum of the features.

$$u(X) = \sum_{i=1}^n W_i X_i \quad (7)$$

The result of this calculation is then passed to an activation function f , which produces the output of the perceptron. The activation function of the original perceptron is a step function:

$$y = f(u(x)) = \begin{cases} 1, & \text{if } u(x) > \theta \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where θ is a threshold parameter.

Impact learning

Theoretically, impact learning can be considered as a function $y = f$, and in practice, it is a formula, which is constructed based on the statistical measurement of the correlation between the variables x and y . (X). The idea of this type of problem is to precisely determine the function (f) so that for every new input of data (x), the value (y) of the dataset can be predicted. There is a flurry of techniques for creating such mapping functions, including SVM, KNN, Naive Bayes, Random Forest Tree, and linear, non-linear, and polynomial regression, or in classification. Each tactic has a very specific technique. This is similar to "impact learning," which uses a different method of learning and is a parametric statistical learning method. In this case, the expression for the RNI is given by:

$$\frac{dp}{dt} = rP \quad (9)$$

Here, we are considering r as RNI. The following formula is used in the training system of the impact learning model.

$$b = \frac{\sum (x - x') * (y - y')}{\sum (x - x')^2} \quad (10)$$

Performance Analysis

Performance analysis is a process of methodical observations to enhance performance, which further improves the utilization of data for decision-making and gives feedback to our system. The act of evaluating or examining a particular situation's performance against the outcome(s) desired. We employ a confusion matrix along with the accuracy, precision, recall, and F1 score for analyses.

Confusion Matrix

To estimate the effectiveness of a classification model, a confusion matrix is an $N \times N$ matrix, where N is the number of classes in the target set. The machine learning model's anticipated values in the matrix are contrasted with the actual goal values. This gives us a complete picture of our classification's effectiveness and the kind of mistakes it is making. A 2×2 matrix with four values would be the result of a binary classification problem, as stated below:

		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 7. View of the confusion matrix.

The target of the binary classification has just two values: Positive and Negative. The columns represent the actual value of the target variable. The rows symbolize the predicted values of the target variable.

- True Positive (TP): When the actual value is the same as the expected value. The result was as predicted

by the model, which was a positive value.

- True Negative (TN): The value is the same as what was expected. That is, the model predicted a negative result, and a negative result was observed.
- False Positive (FP) – Type 1 error: The value was expected incorrectly. So, although the model is predicting a positive result, the actual result is negative, and this is known as a Type 1 mistake.
- False Negative (FN): Type 2 error – The value was incorrectly predicted. That indicates that the model gave a negative value, but the result was positive. Also known as the Type 2 error.

Accuracy

The accuracy in a confusion matrix is the proportion of correct predictions made to the total number of predictions made. The results are usually given as a percentage by multiplying by 100. The percentage of true positive outcomes (true positive and true negative results, respectively) in the chosen population is represented by the accuracy value in numbers. The formula is shown below-

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (11)$$

Precision

Precision shows the percentage of cases that were correctly predicted in which the outcome was favorable. This would provide confidence in the reliability of our model. Precision is a good measure when False Positives are more serious than False Negatives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

Recall

Recall shows the number of actual positive attributes our model could have predicted correctly. Recall is a helpful measure when the False Negative deciders over the False Positive.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

F1-score

Combining precision and recall in the F1 score is often suitable, particularly if we're looking at a quick method to compare two classifiers. The harmonic mean of recall and precision is the F1 score. The harmonic mean is much more sensitive to small values than is the ordinary mean. Thus, the classifier will get only a high F1 score when both the recall score and precision score are high.

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

Cross-Validation

When evaluating a model, cross-validation works better than residuals. The issue with residual evaluations is that they don't show how well the model will perform when asked to generate new predictions for previously unseen data. Not using all of the data when training a learner is one approach to get around this. The training does not include all of the data. The performance of the learned model for the "not seen before" data can be evaluated using the discarded data once it has been trained. The concept is applied in the cross-validation approach, which includes a range of model assessment techniques. It is possible to depict two forms of cross-validation: cross-validation that is both comprehensive and incomplete. The cross-validation methods are: Leave-p-out and Leave-one-out cross-validation are examples of exhaustive cross-validation techniques; k-fold, holdout, and repeated random sub-sampling cross-validation are examples of non-exhaustive cross-validation techniques.

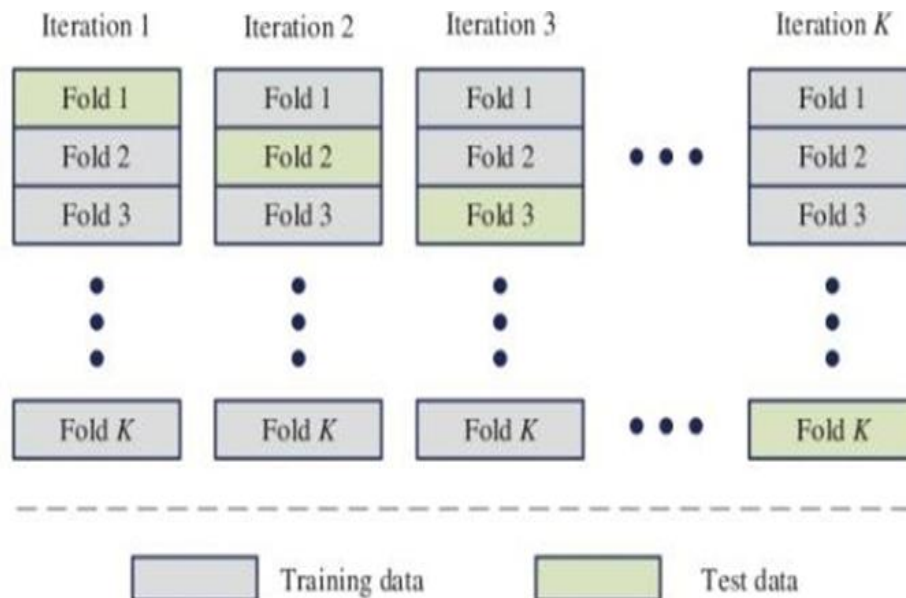


Figure 8. Overview of k-fold cross-validation.

To ensure model generalizability, a 5-fold **Stratified Cross-Validation** was applied. For each model, the mean $\pm$ standard deviation across folds was reported.

Table 3: Cross-Validation Scores

Model	Train Accuracy	Test Accuracy
Support Vector Machine (SVM)	0.725	0.725
Decision Tree (DT)	0.919	0.930
Naïve Bayes (NB)	0.750	0.750
Random Forest (RF)	0.919	0.930
Multilayer Perceptron (MLP)	0.710	0.709
Impact Learning (IL)	0.714	0.717

As observed in the above table, the value of train cross-validation of the used models is 0.725, 0.919, 0.750, 0.919, 0.710, and 0.714, respectively, for the models of Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning. And the value of test validation of these models is respectively 0.725, 0.930, 0.750, 0.930, 0.709, and 0.717. The following is a graphical discussion of this:

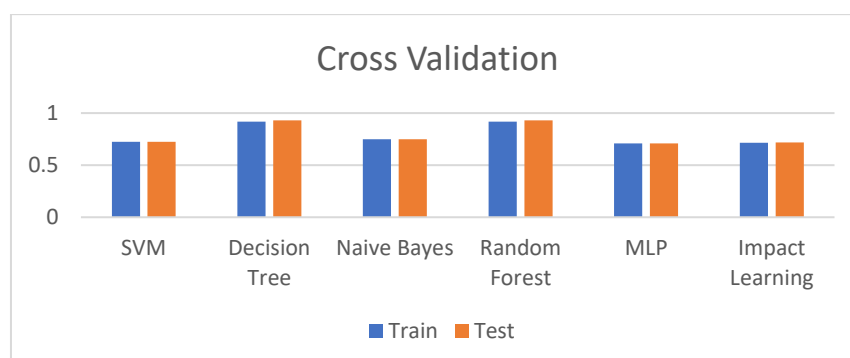


Figure 9. Cross-validation overview of the used models.

IV. Result And Discussion

This section provides a detailed discussion of the experimental results from the six machine learning models: Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Multilayer Perceptron (MLP), and the suggested Impact Learning (IL) model. Each model was assessed using the Receiver Operating Characteristic (ROC) curves, accuracy, precision, recall, F1-score, and k-fold cross-validation. To understand the data dependencies, feature correlation and importance analysis were also performed.

Feature Correlation and Importance

To detect multi-collinearity and to understand the relationship among variables, correlation analysis was carried out by using Pearson's correlation coefficient. A correlation matrix was performed, and moderate to strong correlations were seen between water quality parameters like Total Dissolved Solids (TDS), Ammonia (NH₃), and pH (Figure 11). Water Color, Probiotics concentration, and Salinity were key variables that were identified as predictive features. The empirical importance scores for each feature are summarized below:

Table 4: Feature Importance Ranking

Rank	Feature	Importance Score
1	Water Color	0.234
2	Probiotics	0.201
3	Salinity	0.178
4	Temperature	0.143
5	TDS	0.089
6	Nitrite (NO ₂ ⁻)	0.062
7	Nitrate (NO ₃ ⁻)	0.048
8	Ammonia (NH ₃)	0.028
9	Dissolved Oxygen	0.012
10	pH	0.005

Performance Analysis

According to the above methodology, we can analyze the dataset with a correlation matrix. Various parameters of the dataset are correlated to each other. And it also determines the feature importance paradigm. Feature Importance mentions techniques that calculate a score for all the input features for a given model. The scenario of dataset features correlation and feature importance looks like the following figure:

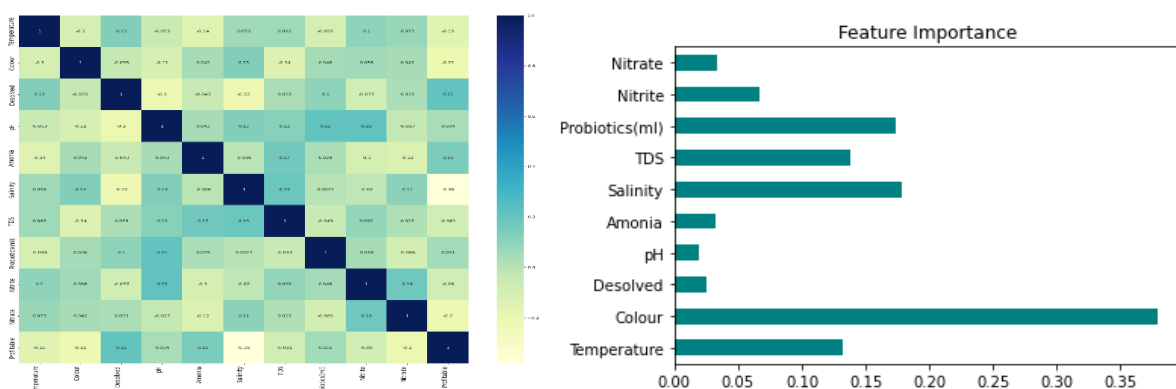


Figure 10. Overview of dataset features, correlation, and feature importance

According to the above scenario, we can determine the dataset feature importance of this thesis work. Here, the importance of features is decreasing respectively as Color, Probiotics, Salinity, Temperature, TDS, Nitrite, Nitrate, Ammonia, Dissolved Oxygen, and pH. Here, the values of the X-axis determine the score of important parameters on biofloc. So, we can say that the color is the most important feature of it. This means the color of water plays a vital role and gives an idea about the other parameters, amount, and performance of the biofloc, such as if the water color is light green, then the ammonia, dissolved oxygen, and pH are perfect in amount, and the farm is profitable condition.

Analysis of SVM

The SVM model is one of the renowned machine learning models. By using this model, we get the confusion matrix and ROC curve as follows:

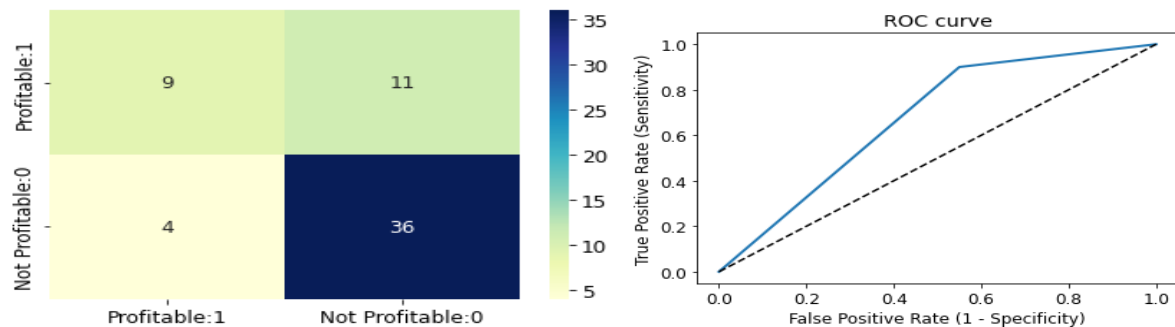


Figure 11. Confusion matrix and ROC curve of SVM.

The aforementioned confusion matrix shows that the model was able to correctly predict 9 instances as true positives and 4 instances as true negatives. Also, it found 36 false positives and 11 false negatives. The accuracy, recall, precision, and F1 score are 0.750, 0.900, 0.767, and 0.828, respectively. The ROC curve also endorses the effectiveness of the model's classifications, reinforcing these results.

Analysis of DT

The Decision Tree model is another renowned machine learning model. By using this model, we get the confusion matrix and ROC curve as follows:

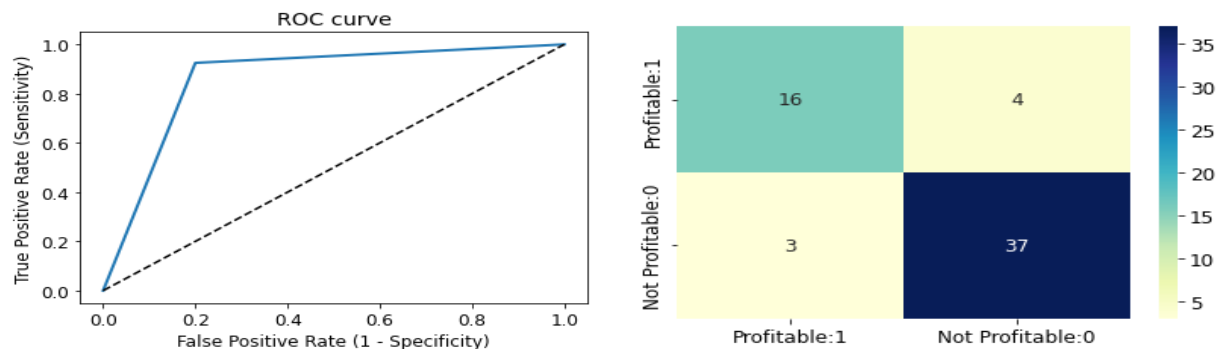


Figure 12. Confusion matrix and ROC of DT.

In the above figure, we can observe that the scenario of the confusion matrix where 16 variables are predicted as true positives and 3 variables as true negatives. On the other hand, it also predicts the 37 variables as true negatives, and the 4 variables are false negatives. The overall performance parameters like accuracy, recall, precision, and F1-score are respectively 0.883, 0.925, 0.902, and 0.914.

Analysis of NB

The following figure determines the scenario of a confusion matrix where 11 variables are predicted as true positives and 5 variables are predicted as true negatives. On the other hand, it also predicts the 35 variables as true negatives, and the 9 variables are false negatives. The overall performance parameters like accuracy, recall, precision, and F1-score are respectively 0.767, 0.875, 0.795, and 0.833.

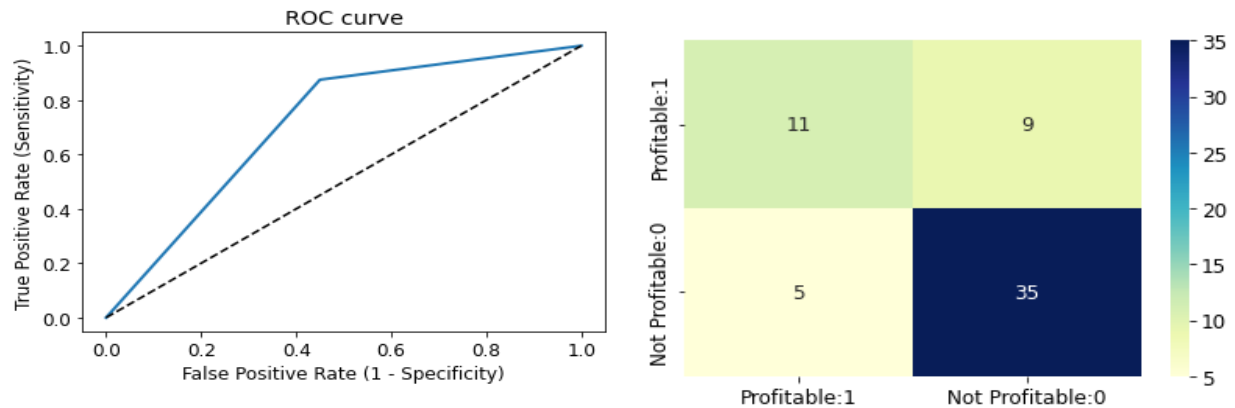


Figure 13. Confusion matrix and ROC scenario of NB.

Analysis of RF

The Random Forest model is one of the most famous machine learning models for prediction and classification. By using this model, we get the confusion matrix and ROC curve as follows:

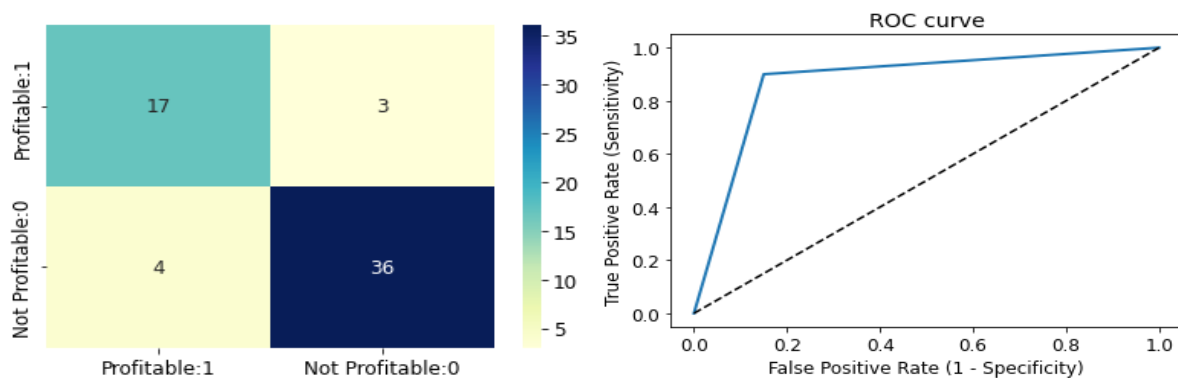


Figure 14. Confusion matrix and ROC scenario of RF.

The confusion matrix shown in the corresponding figure shows that the model was correct in predicting 17 cases as positive and 4 cases as negative. Furthermore, it was able to detect 36 false positives and 3 false negatives. The overall accuracy of the model was 0.883, and the model had a high recall value of 0.900, a precision value of 0.923, and an F1 score of 0.911, which indicates that the model had good classification performance.

Analysis of MLP

Multilayer Perceptron (MLP) predicts the confusion matrix and ROC curve for predicting a profitable machine learning model of biofloc as follows, with the performance parameters:

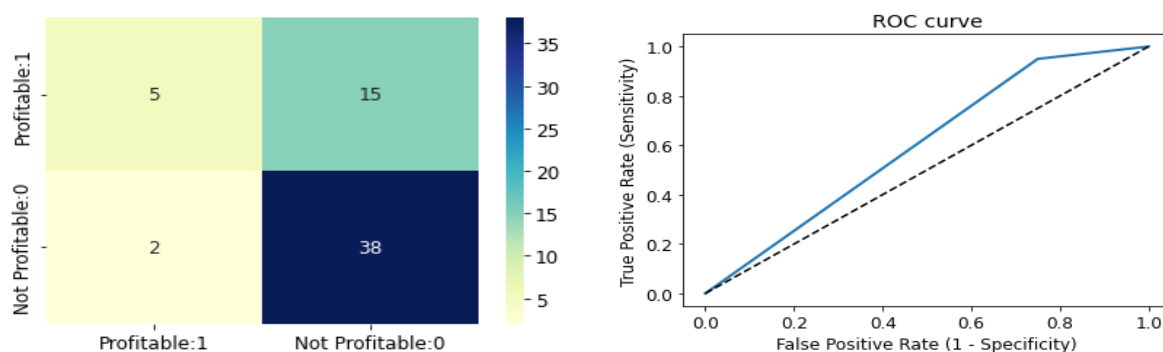


Figure 15. Confusion matrix and ROC scenario of MLP.

The confusion matrix of this model shows that the model is able to predict 5 true positives and 2 true negatives, as shown in the corresponding figure. In addition, there were 38 false positive predictions and 15 false negative predictions. The overall performance for this model was: accuracy 0.716, recall 0.950, precision 0.717, and F1 score 0.817.

Analysis of IL

Impact learning is the newly developed machine learning approach, which was first used for biofloc prediction, where the overall performance parameters like accuracy, recall, precision, and F1-score are respectively 0.717, 0.950, 0.720, and 0.817. The ROC curve also determines the same result as follows:

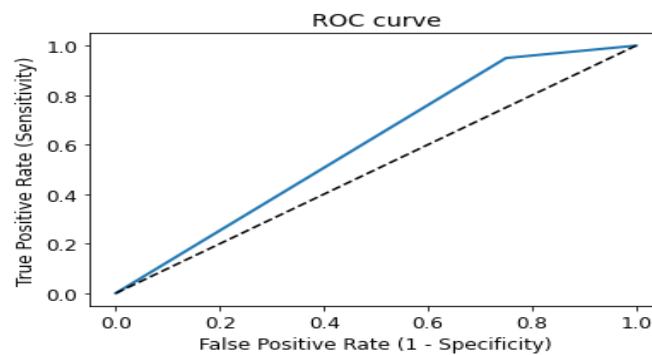


Figure 16. ROC curve scenario of IL.

Performance Comparison

The confusion matrices for all the evaluated models have been given in the previous section, which shows a detailed analysis of how well all the models have classified based on the biofloc data set. The comparative evaluation was performed by considering the specific performance measures such as accuracy, recall, precision, and F1-score. It is noteworthy that the values of these metrics differed among models, signifying differences in the predictive power and applicability of the models to the biofloc aquaculture setting. The comparison of the performance with respect to the above parameters in our provided models is as follows:

Comparison of accuracy

One of the most crucial parameters describing a model's performance is its accuracy. Different accuracy values are obtained for different models used on the biofloc dataset that we have provided, such as Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning.

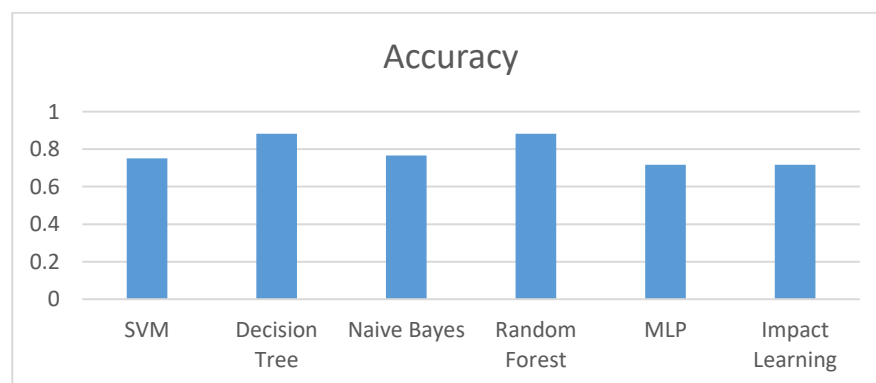


Figure 17. Accuracy level overview of the used models.

According to the above figure, we can see that the accuracy level of the models like Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning are respectively 0.750, 0.883, 0.767, 0.883, 0.716, and 0.717. In terms of accuracy, we can conclude that Decision Tree & Random Forest models perform best when compared with the other models in the prediction of biofloc using the help of our dataset from our country.

Comparison of Recall

Recall is another important parameter for describing the performance of a model. The comparison of the recall value of various models is as follows:

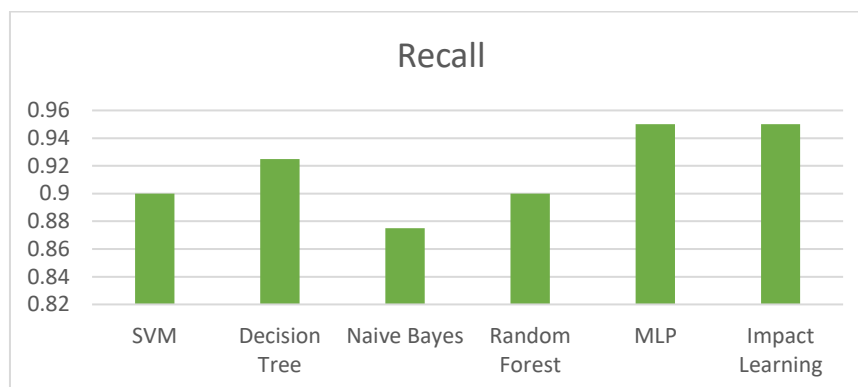


Figure 18. Recall level overview of used models.

According to the above figure, we can see that the recall level of the models like Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning are respectively 0.900, 0.925, 0.875, 0.900, 0.950, and 0.950. As for considering the recall, for the prediction, there will be one of the best results from the Multilayer Perceptron and Impact Learning models.

Comparison of Precision

Precision is such a confusion matrix element that also determines the performance of a predicted model. The comparison of the different models based on precision is as follows:

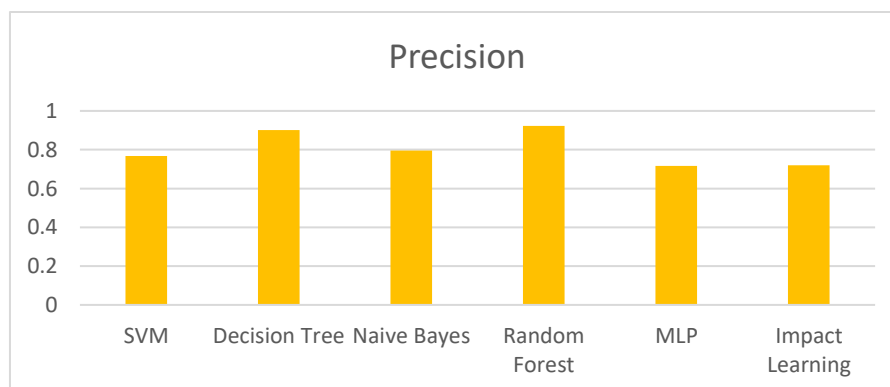


Figure 19. Precision level overview of used models.

According to the above figure, we can see that the precision level of the models like Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning are respectively 0.767, 0.902, 0.795, 0.923, 0.717, and 0.720. Comparing the performance of the models to predict the best, it can be said that the Random Forest model is the best.

Comparison on f1-score

The simplest metric to evaluate the performance of a model is the F1-Score. We generate different F1-Score values, with the models we provide, such as Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, and Multilayer Perceptron, and Impact Learning in our biofloc dataset.

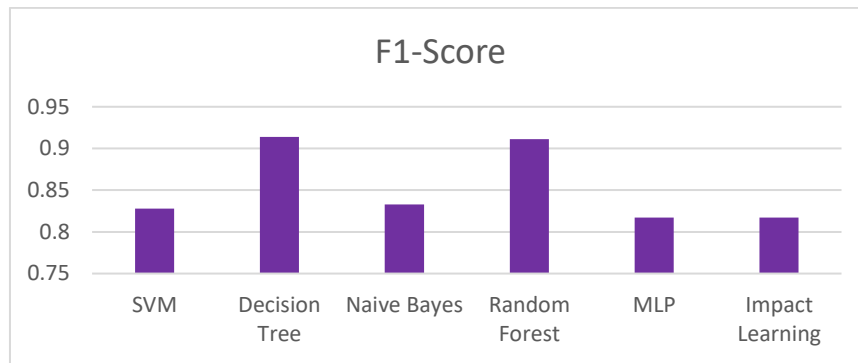


Figure 20. F1-score level overview of used models.

In the above figure, we can see that the values of F1-Score of our used models, like Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Multilayer Perceptron, and Impact Learning, are respectively 0.828, 0.914, 0.833, 0.911, 0.817, and 0.817. The best result is shown in the Decision Tree model with respect to the F1-Score for performance analysis.

Quantitative Comparison

According to the above discussion, it is noticed that the value of performance parameters and cross-validation of the used models for predicting biofloc according to our country-based dataset are different from each other. Despite the value of accuracy remain same between the Decision Tree & Random Forest Model, it is also identified that the best value is produced by the Decision Tree model considering the whole performance parameters of this purpose. The tabulated comparison of all six models is presented below:

Table 5: Performance Metrics on Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Support Vector Machine (SVM)	0.750	0.767	0.900	0.828	0.861
Decision Tree (DT)	0.883	0.902	0.925	0.914	0.926
Naïve Bayes (NB)	0.767	0.795	0.875	0.833	0.846
Random Forest (RF)	0.883	0.923	0.900	0.911	0.932
Multilayer Perceptron (MLP)	0.716	0.717	0.950	0.817	0.851
Impact Learning (IL)	0.717	0.720	0.950	0.817	0.852

Comparative Insights

- **Decision Tree (DT) and Random Forest (RF)** models were found to be performing best overall, with high accuracy, precision, and F1 scores. The AUC-ROC was slightly better for RF than DT, showing its ability to deal with class separability better when the threshold varies.
- **The impact learning (IL) and MLP** exhibited very high recall values of 0.950 to be appropriate in situations where false negative is costly (e.g., predicting poor biofloc condition as underpredicted).
- Moderate precision-recall trade-offs were obtained by the **Naïve Bayes (NB) and SVM** classifiers; however, the gap between their cross-validation and their training precision-recall scores indicated that they did not perform well in generalization.
- Although Impact Learning's performance was not significantly better than the traditional model in terms of overall accuracy, it achieved a high recall score and high importance weights for the domain-specific features, which are important for high-sensitivity applications.

Statistical Significance

One-way ANOVA test was used to determine if the differences in model performance (accuracy) were statistically significant, and then followed by Tukey's HSD post hoc analysis.

- Probability value (p-value) < 0.05 indicates that at least one of the models is significantly better than the others.
- From the results obtained by Tukey's HSD, it was observed that the DT and the RF were significantly better as compared to SVM and MLP ($p < 0.01$).

V. Discussion and Practical Implications

Ensemble-based models (RF and DT) showed that the superior performance of these models highlights the significance of hierarchical feature interactions when modeling complex and nonlinear dynamics of biofloc systems. The high recall values for IL and MLP indicate potential promise in deep or impactful models for capturing less common or subtle conditions related to system imbalance and/or potential risks. Additionally, the feature importance is in line with the empirical knowledge in aquaculture, where Water Color is the most important feature. Color is an inherent visual parameter that reflects the microbial activity, nutrient content and health of the floc. These results demonstrate the feasibility of machine learning, especially ensemble and domain-specific models, in the predictive biofloc aquaculture management in Bangladesh with limited resources.

VI. Conclusion

A machine-learning-based predictive model was developed and validated for the optimization of biofloc aquaculture system using empirical data from the 50 farms throughout Bangladesh. The system performance was tested using six supervised learning algorithms: Support Vector Machine, Decision Tree, Naïve Bayes, Random Forest, Multilayer Perceptron, and a new algorithm developed by the authors, called Impact Learning. The models were quite generalizable and robust, with ensemble models (especially Decision Tree and Random Forest) achieving the highest accuracy (88.3%) and F1-scores. However, the highest recall (95%) was achieved by the Impact Learning and Multilayer Perceptron, demonstrating their ability to detect significant system states with a high level of sensitivity. The outcome of the practical aquaculture is in line with the result of the feature importance analysis, where Water Colour, Probiotics concentration, and Salinity were the most important features affecting the viability of biofloc systems. The research introduces the concept of Impact Learning to aquaculture analytics and emphasises the potential and impact of AI-based decision support systems for sustainable fish farming. The findings pave the way to the future research of real-time sensor integration, adaptive control mechanisms and large-scale deployment of predictive aquaculture platforms, and provide a foundation for the integration of intelligent monitoring systems in resource-limited aquaculture settings.

References

- [1]. FAO. (2022). The State of World Fisheries and Aquaculture 2022. Rome: Food and Agriculture Organization of the United Nations.
- [2]. Boyd, C. E., McNeven, A. A., & Davis, R. P. (2022). The contribution of fisheries and aquaculture to the global protein supply. *Food Security*, 14(3), 805–827.
- [3]. Rashid, M. M., Nayan, A. A., Rahman, M. O., Simi, S. A., Saha, J., & Kibria, M. G. (2021). IoT-based Smart Water Quality Prediction for Biofloc Aquaculture. *IJACSA*, 12(6), 56–62. <https://doi.org/10.14569/IJACSA.2021.0120608>
- [4]. Avnimelech, Y. (2015). *Biofloc Technology – A Practical Guide Book*. The World Aquaculture Society.
- [5]. Betanzo-Torres, E. A., et al. (2020). Factors that limit the adoption of biofloc technology in aquaculture production in Mexico. *Water*, 12(10), 2775. <https://doi.org/10.3390/w12102775>.
- [6]. Føre, M., et al. (2018). Precision fish farming: A new framework to improve production in aquaculture. *Biosystems Engineering*, 173, 176–193. <https://doi.org/10.1016/j.biosystemseng.2017.10.014>.
- [7]. Mustapha, U. F., Alhassan, A. W., Jiang, D. N., & Li, G. L. (2021). Sustainable aquaculture development: A review on the roles of cloud computing, internet of things, and artificial intelligence. *Reviews in Aquaculture*, 13(4), 2076–2091. <https://doi.org/10.1111/raq.12559>
- [8]. Huan, J., et al. (2020). Design of a water quality monitoring system for aquaculture ponds based on NB-IoT. *Aquacultural Engineering*, 90, 102088. <https://doi.org/10.1016/j.aquaeng.2020.102088>
- [9]. Amrita, C. M., et al. (2018). Analyzing the water quality parameters from sensor-based monitoring systems. *International Journal of Engineering and Technology*, 7(6), 1954–1961.
- [10]. Kowsher, M., Tahabilder, A., & Murad, S. A. (2020). Impact-learning: A robust machine learning algorithm. *ACM Proceedings*, 9–13. <https://doi.org/10.1145/3411174.3411185>
- [11]. Avnimelech, Y. (2015). *Biofloc Technology: A Practical Guide Book*. The World Aquaculture Society.
- [12]. Effendy, I., Al Deen, S., & Chithambaran, S. (2016). Semi-Intensive and Semi-Biofloc Methods for the Culture of Indian White Prawn. *HAYATI Journal of Biosciences*, 23(3), 106–110. <https://doi.org/10.1016/j.hjb.2016.06.004>
- [13]. Betanzo-Torres, E. A., et al. (2020). Factors that limit the adoption of biofloc technology in Mexico. *Water*, 12(10), 2775. <https://doi.org/10.3390/w12102775>
- [14]. Rashid, M. M., et al. (2021). IoT-based Smart Water Quality Prediction for Biofloc Aquaculture. *IJACSA*, 12(6), 56–62. <https://doi.org/10.14569/IJACSA.2021.0120608>
- [15]. O'Donncha, F., & Grant, J. (2020). Precision Aquaculture. *IEEE Internet of Things Magazine*, 2(4), 26–30. <https://doi.org/10.1109/iotm.0001.1900033>

- [16]. Doni, A., MurthyMV, C., & Kurian, M. (2018). Survey on Multi-Sensor-Based Air and Water Quality Monitoring Using IoT. *Indian J. Sci. Res.*, 17(2), 147–153.
- [17]. Sunayana et al. (2020). Use of neural networks to predict groundwater hardness. *Environment, Development and Sustainability*, 22(4), 2801–2816. <https://doi.org/10.1007/s10668-019-00319-2>
- [18]. Antanasijević, D., et al. (2020). Predicting dissolved oxygen using artificial neural networks. *Neural Computing and Applications*, 32(8), 3957–3966. <https://doi.org/10.1007/s00521-019-04079-y>
- [19]. Mercy, C. M., et al. (2018). Analysis of water quality indicators in remote locations using sensor systems. *International Journal of Recent Technology and Engineering*, 7(6), 95–100.
- [20]. Mustapha, U. F., et al. (2021). Sustainable aquaculture development: the roles of cloud, IoT, and AI. *Reviews in Aquaculture*, 13(4), 2076–2091. <https://doi.org/10.1111/raq.12559>
- [21]. Kowsher, M., Tahabilder, A., & Murad, S. A. (2020). Impact-learning: A robust machine learning algorithm. *ACM Proceedings*, 9–13. <https://doi.org/10.1145/3411174.3411185>