Original article

Evaluation of disease outbreak in terms of physico-chemical characteristics and heavy metal load of water in a fish farm with machine learning techniques

Mesut Yilmaz [a], Mustafa Çakir [b], Mükerrem Atalay Oral [c],*, Hüseyin Özgür Kazanci [d], Okan Oral [d]

[a] Akdeniz University, Faculty of Fisheries, Antalya, Türkiye
[b] İskenderun Technical University, Iskenderun Vocational School of Higher Education, İskenderun, Hatay, Türkiye
[c] Akdeniz University, Elmalı Vocational School of Higher Education, Antalya, Türkiye
[d] Akdeniz University, Faculty of Engineering, Antalya, Türkiye

ARTICLE INFO

ABSTRACT

Diseases are quite common in fish farms because of changes in physico-chemical characteristics in the aquatic environment, and operational concerns, i.e., overstocking and feeding issues. In the present study, potential factors (water physico-chemical characteristics and heavy metal load) on the disease-causing state of the pathogenic bacteria *Lactococcus garvieae* and *Vagococcus* sp. were examined with machine learning techniques in a trout farm. Recording of physico-chemical characteristics of the water, fish sampling and bacteria identification were carried out at bimonthly intervals. A dataset was generated from the physico-chemical characteristics of the water and the occurrence of bacteria in the trout samples. The eXtreme Gradient Boosting (XGBoost) algorithm was used to determine the most important independent variables within the generated dataset. The most important seven features affecting bacteria occurrence were determined. The model creation process continued with these seven features. Three well-known machine learning techniques (Support Vector Machine, Logistic Regression and Naïve Bayes) were used to model the dataset. Consequently, all the three models have produced comparable results, and Support Vector Machine (93.3% accuracy) had the highest accuracy. Monitoring changes in the aquaculture environment and detecting situations causing significant losses through machine learning techniques have a great potential to support sustainable production.

1. Introduction

Disease outbreaks commonly occur in fish farms. Diseases persist despite the invention of vaccinations and novel culturing techniques (Kusuda and Kawai 1998). Bacterial infections tend to arise quite frequently because of changes in physico-chemical characteristic of the water, such as pH, dissolved oxygen concentration,

and presence of iron in water environment, as well as operational issues, including overstocking and feeding issues (Hansen and Olafsen 1999; Verschuere et al. 2000; Winton 2001). They can result in significant economic losses such as fish fatalities and growth stalling if left untreated.

Bacterial disease caused by *Lactococcus garvieae* is one of the most prevalent diseases in rainbow trout farming and can result in significant financial losses. Lactococcus infections, which affect fish of all sizes and are particularly prevalent in the summer and prevent breeding (Diler et al., 2002). Vagococcosis is another disease that can be seen in rainbow trout farms. This disease rarely emerges but affects large numbers of fish at 10–12 °C (Ruiz-Zarzuela et al., 2005).

Metals are naturally occurring elements of the aquatic systems (Nieboer and Richardson, 1980). While some non-essential metals, i.e., cadmium (Cd) and lead (Pb) have no biological function in animals and can still have hazardous effects even at low concentrations, the others, i.e., copper (Cu) and chromium (Cr) are

Production and hosting by Elsevier

necessary for biochemical and physiological processes in the majority of organisms (Bryan and Darracott 1979; Tchounwou et al., 2012; Williams, 1996). The immunological competence of fish is affected by exposure to several heavy metals, including Cd, Cr, Cu, and Pb. Stressors like metals may alter fish susceptibility to diseases brought on by pathogenic bacteria, fungi, and virus since proper cellular and humoral immune responses are required for protection against diseases in fish (Zelikoff, 1993).

The application areas of the techniques of machine learning (ML), which is one of the sub-fields of artificial intelligence are increasing day by day (Liakos et al., 2018). One of these areas is the determination of the factors that play an active role in the detection of diseases encountered in farm trout farming. Investigation of the factors that cause disease in trout farming is important in terms of production quality and efficiency. For this reason, both the physico-chemical characteristic of the farm water and the effects of heavy metals in the water are investigated in farms.

In the present study, the effects of physico-chemical characteristics and heavy metal content of water on disease outbreak caused by pathogenic bacteria (*Lactococcus garvieae* and *Vagococcus* sp.) on rainbow trout were investigated. In this context, physico-chemical characteristics, and toxic metal content of water at the inlet and outlet of the farm were determined. Pathogenic bacteria profile was monitored at the farm. Among all features, the most important features were obtained with the XGBoost algorithm, which is one of the most popular ML techniques of the recent period. The models of the obtained data were created with three different widely used ML techniques, including logistic regression (LR), support vector machine (SVM) and Naïve bayes (NB). Finally, seven different metrics were used to evaluate the models used for classification. Finally, it was tested whether the disease outbreak could be previously predicted via features by using ML techniques or not.

2. Materials and methods

2.1. Overview of the study area

The present study was carried out in a fish farm located in Mugla province (Türkiye). The water utilized in the farm is directly taken from a spring, and there are no other farms affecting the water quality in that area. Furthermore, there are very few seasonal variations in the pH and temperature of inlet water. For these reasons, this farm was selected as a study area. The physico-chemical characteristics of the water of the aquaculture environment were measured by using a multiparameter instrument (YSI, USA). In fish sampling and bacteria identification, a total 6 samplings were conducted at bimonthly intervals. Ten specimens were taken at each sampling (a total of 60 fish). The bacteria were diagnosed at the Fish Disease Laboratory at Eğirdir Faculty of Fisheries (Isparta University of Applied Sciences).

2.2. Description of the collected data

Physico-chemical and heavy metal measurements were conducted from the farm inlet (1) and farm outlet (2) water to determine the possible factors that may affect the emergence of *Lactococcus garvieae* and *Vagococcus* sp. bacteria. The physico-chemical characteristics of water were temperature (T), pH value (pH), saturation (Sat), dissolved oxygen (DO), and total dissolved solids (TDS). In addition, measurements were performed for heavy metals such as Cr, Cu, Cd, and Pb. Explanations of these features in the data set are given in Table 1.

The XGBoost algorithm was used for feature selection (FS) in the current study. The XGBoost is used for an efficient conduct of the gradient boosting algorithm (Friedman, 2001). The XGBoost algorithm can maximize the prediction performance (Alshari et al., 2021). Despite the success and popularity of XGBoost among ML techniques, no scientific studies have been found examining the effects of toxic metals and physico-chemical characteristics of water on the emergence of pathogenic bacteria in trout farms by using this algorithm.

Since the infected or healthy fish indicate two different classes, binary classification methods from ML techniques were used to solve the problem in this study. According to Binkhonain and Zhao (2019) and Shehab et al. (2022), the most popular supervised learning classification techniques are SVM, NB and LR. For this rea-son, classification models were created with the mentioned ML techniques in this study.

Polynomial regression or logistic regression, which started to be used in the science of biology at the beginning of the 20th century, has been used in many fields of science today (Önder and Cebeci, 2014). Advantages such as being easy to interpret, applying to large data sets and not requiring assumptions have made model building with LR popular among ML tools (Vupa Çilengiroğlu and Yavuz, 2020). For these reasons, SVM (a versatile and popular supervised ML algorithm that can be used for regression and classification)

Table 1
Descriptions of features in the data set obtained from the fish farm.

No.	Feature Name	Description	Type
1	T1	Inlet water temperature (°C)	Numeric independent
2	pH1	Potential of hydrogen value of inlet water	Numeric independent
3	Sat1	Inlet water saturation (%)	Numeric independent
4	DO1	Dissolved O_2 concentration (mg/L) of inlet water	Numeric independent
5	TDS1	Total dissolved solids (g/L) in inlet water	Numeric independent
6	T2	Outlet water temperature (°C)	Numeric independent
7	pH2	Potential of hydrogen value in outlet water	Numeric independent
8	Sat2	Outlet water saturation (%)	Numeric independent
9	DO2	Dissolved O_2 concentration (mg/L) of outlet water	Numeric independent
10	TDS2	Total dissolved solids (g/L) in outlet water	Numeric independent
11	Cr1	Inlet water Chromium concentration (μg/L)	Numeric independent
12	Cr2	Chromium concentration (μg/L) in outlet water	Numeric independent
13	Cu1	Inlet water Copper concentration (μg/L)	Numeric independent
14	Cu2	Copper concentration (μg/L) in outlet water	Numeric independent
15	Cd1	Inlet water Cadmium concentration (μg/L)	Numeric independent
16	Cd2	Cadmium concentration (μg/L) in outlet water	Numeric independent
17	Pb1	Inlet water Lead concentration (μg/L)	Numeric independent
18	Pb2	Lead concentration (μg/L) in outlet water	Numeric independent
19	Month	Sampling Month	Categorical independent
20	Bacteria Occurrence	Detected (1) or non-detected (0) of bacteria (*Lactococcus garvieae* and *Vagococcus sp.*)	Categorical dependent

and NB (Rish, 2001) (a statistical classification method using the classifier and similarity attribute) were used in this study. Although NB is a simple method, it is preferred in classification problems in many areas.

2.3. ML techniques

The ML techniques, which are a sub-field of AI are divided into two groups. These are supervised and unsupervised learning. Supervised learning, unlike unsupervised learning, makes predictions or classifications for new data by using labeled data (Kumar et al., 2020). Regression, classification, and clustering operations can be performed with supervised learning. The classification process is also called binary classification if there are two classes, and multiple classifications if there are more than two classes (Prabhakaran, 2016). Because whether the fish is infected or healthy indicates two different classes, binary classification methods were used to solve the problem in the present study. According to Binkhonain and Zhao (2019) and Shehab et al. (2022) the most popular supervised learning classification techniques are SVM, NB and LR. For this reason, classification models were created with the mentioned ML techniques in this study.

2.3.1. Logistic regression
Polynomial regression, or logistic regression has been used in many fields of science today (Önder and Cebeci, 2014). Advantages such as being easy to interpret, applying to large data sets and not requiring assumptions have made model building with LR popular among ML tools (Vupa Çilengiroğlu and Yavuz, 2020).

The LR can be used if the target variable, i.e., dependent variable is categorical. In LR, the aim is to establish a linear model that describes the relationship between dependent and independent variables. To create the model, it is necessary to first look at the logistic distribution function. The linearization of the logistic distribution function is possible with the logit transform. Otherwise, linear regression will give results in the range (-∞,+∞) (James et al., 2021).

The logistic distribution function is included in Eq. (1).

$$\pi(x) = P(Y = 1 \mid X = x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}} \tag{1}$$

In the equation, X = independent variables, x = the values that the independent variables can take, β= regression coefficients, u= error intercept, p = the number of X independent variables in the dataset, Y = class value, and $\pi(x)$ = the output value of the dependent variable in the range of 0–1.

The $\pi(x)$ function gives the probability of an event occurring, while the 1- $\pi(x)$ function gives the probability of an event not happening. Therefore, the ratio of the probability of an event occurring to the probability of not occurring in the logistic regression function is as in Eq. (2).

$$\frac{\pi(x)}{1 - \pi(x)} \tag{2}$$

When the natural logarithm of the proportioning process is taken, the LR model is obtained. The equation of the model is as in Eq. (3).

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3}$$

The LR calculates the probability of occurrence of the value defined as 1 for the dependent variable. For this reason, it does not care about the value of the dependent variable. Although the model can be constructed simply in LR, the main challenge is to optimize the model with higher accuracy (Baydemir, 2021).

2.3.2. Support vector Machine
SVM is a versatile and popular supervised ML algorithm that can be used for regression and classification. SVM is generally preferred for small and medium-sized classification problems. SVM basically finds a hyperplane in N-dimensional space that distinctly separates data points belonging to classes. A hyperplane is a plane that linearly divides n-dimensional data into two sections. A hyperplane is a line in two dimensions, a plane in three dimensions. The hyperplane is the line that can linearly separate the data of classes 1 and 2. SVM for binary classification was developed by (Cotes and Vapnik, 1995). The mathematical expression for a 2-dimensional hyperplane is seen in Eq. (4).

$$y = a * x + b \tag{4}$$

To convert the mathematical expression of the hyperplane to vector, Equation (5) was used.

$$a * x + b - y = 0 \tag{5}$$

Let X = (x,y) vector and W = (a,-1), then from hyperplane to vector with Equation (6).

$$W.X + b = 0 \tag{6}$$

Assuming that the binary classification training data with SVM are labeled as $\{x_i, y_i\}$, $i = 1 \cdots l$, $y_i \in \{-1, 1\}$, $x_i \in \Re^d$ in a two−dimensional space will be $x_i \in \Re^2$. It is assumed that the data can be separated linearly and there are many hyperplanes in the separation (Prabhakaran, 2016).

SVM (Cotes and Vapnik, 1995) solves the constrained frame optimization problem algorithmically, creating an optimum separation hyperplane f (x) = 0 between the data sets. The decision function is seen in Eq. (7).

$$y = f(x) = xW^T + b = \sum_{i=1}^{k} (W_i x_i + b) \tag{7}$$

2.3.3. Naïve Bayes
NB classifier is a statistical classification method using similarity attribute (Rish, 2001). Although it is a simple method, it is preferred in classification problems in many fields. The NB classifier is expressed by Bayes' theorem (Berrar, 2018). The equation for the theorem is as in Eq. (8).

$$P(H|X) * P(X) = P(X|H) * P(H) \tag{8}$$

Here, the conditional probability of the event H to the event X is defined as P (H|X). In other word, it is the probability of the occurrence of the event H when the event X occurs. P(H) and P(X) express the probability of the occurrence of the events H and X (Taheri and Mammadov, 2013).

$X = [X(1), X(2), \cdots X(L)]^T \in R^L$, L-dimensional quantity vector is created. According to the Bayes decision theory, a given X belongs to which class; $P(S_i|X) > P(S_j|X)$, $\forall j \neq i$ belongs to the class S_i.

2.4. Tune of hyper parameters

There are several challenges in ML. Two of the most important of these challenges are deciding which algorithm to choose for model building and how to optimize that algorithm. It is necessary to fine-tune the hyper parameters to create a successful model with the ML technique chosen. This process is called hyper parameter optimization. The hyper parameters of the relevant ML technique should be known before starting the hyper parameter optimization. Some ML techniques may have more than one hyper parameter, while others may have only one or none. The hyper parameters given in a certain range of values are respectively applied to create the model. Accuracy and Kappa metrics are obtained with the method called cross-validation of the model.

After selecting the hyper parameter with the best of these metric values, the hyper parameter optimization is completed. Sometimes, instead of the range to be entered by the user, the search algorithm (grid search, random search, etc.) automatically selects from the search space and produces a solution.

2.5. Model evaluation

Accuracy metric is often used to measure the success of the models created using ML techniques (Alan, 2020). When the Accuracy metric is evaluated without other metrics, it leads to poor generalization results of the classifiers. For this reason, the accuracy metric should be evaluated together with other metrics (Yin et al., 2019). The error matrix used to calculate the accuracy metric value in ML classification problems is the confusion matrix. Confusion matrix is a matrix that is the same size as the number of classes and shows the relationship between actual values and predicted values.

During disease classification, fish can be classified as true positive (TP) or true negative (TN) if correctly diagnosed, false positive (FP) or false negative (FN) if misdiagnosed. The TP, TN, FP and FN values are placed in the confusion matrix. The confusion matrix that can be used to classify a fish as healthy or infected is given in Table 2.

The metrics obtained from the confusion matrix and frequently used to measure binary classification performance are given between Equation (9) and Equation (17) (Hossin and Sulaiman, 2015; Powers, 2011; Prabhakaran, 2016).

Some of the metrics whose values in the confusion matrix are used in the evaluation of classification algorithms are sensitivity, specificity, precision, prevalence, accuracy, F1 score, expected accuracy (EA), and Kappa. Performance metrics for the evaluation of ML techniques, including accuracy, precision, sensitivity, and F1 score, are widely employed in disease diagnosis (Ahsan et al., 2022).

$$N = \text{Total observation count} = TP + FN + FP + TN \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Specificity = \frac{TN}{FP + TN} \quad (11)$$

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

$$Prevalence\,(Pre) = \frac{TP + FN}{N} \quad (13)$$

$$Accuracy\,(Acc) = \frac{TP + TN}{N} \quad (14)$$

$$F1\,Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (15)$$

Table 2
General structure of the confusion matrix.

Binary Classification		Actual	
		Healthy	**Infected**
Predict	**Healthy**	TP	FP
	Infected	FN	TN

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

Table 3
Interpretation of Kappa metric.

Kappa value	Level of agreement	Reliable data (%)
0 – 0.20	None	0–3%
0.21 – 0.39	Minimal	4–15%
0.40 – 0.59	Weak	16–34%
0.60 – 0.79	Moderate	35–63%
0.80 – 0.90	Strong	64–81%
0.91 – 1.00	Almost perfect	82–100%

$$Expected\,Accuracy\,(EA) = \frac{(TP + FP) * (TP + FN) + (FP + TN) * (FN + TN)}{N^2} \quad (16)$$

$$Kappa = \frac{Acc - EA}{1 - EA} \quad (17)$$

Accuracy metric means the total number of correctly identified samples across all the observations. Precision metric is measured as the ratio of accurately predicted to all the expected positive observations. Sensitivity metric refers only to the true positive measurement, considering the total observations. Specificity metric calculates how many true negatives are properly defined. F1 score metric is the mean of accuracy and sensitivity in a harmonic manner (Hossin and Sulaiman, 2015).

Kappa can take values between −1 and + 1. When it takes the value of −1, it shows that the discord between the two observers is perfect, and when it takes the value of + 1, it shows that the agreement between the two observers is perfect. When it takes a value of 0, it indicates that the agreement between these two observers may be coincidental. Kappa value between 0 and 1 can be interpreted as indicated in Table 3 when proceeding with 0.2 steps (McHugh, 2012).

3. Results

Among the pathogenic bacteria, only *Lactococcus garvieae* and *Vagococcus* sp. were isolated and identified from fish samples collected in August. The physico-chemical characteristics values and the toxic element concentrations (mean, minimum and maximum) of the water environment were evaluated (Fig. 1). The mean inlet water temperature and pH level (minimum–maximum) were 8.1 °C (7.7–8.5 °C) and 8.3 °C (7.85–8.90), respectively. These values demonstrate that the pH and temperature of the inlet water are constant throughout the year. The ratio of total dissolved solids varied between 0.131 and 0.157 g/L (mean 0.143 g/L). The saturation level and dissolved oxygen concentration, however, were measured as 65–111% (mean 92.91%) and 7.09–13.81 mg/L (mean 11.09 mg/L), respectively. The dissolved oxygen concentration and saturation values for the inlet water changed considerably throughout the year.

The mean temperature and pH level of the outlet water were 9.0 °C (7.7–10.9 °C) and 8.0 (7.59–8.47), respectively. The ratio of total dissolved solids was found to be between 0.144 and 0.160 g/L (mean 0.150 g/L). The minimum–maximum rates for the concentration of dissolved oxygen and saturation value were 6.45–11.73 mg/L (mean 9.78 mg/L) and 53–99% (mean 83.50%), respectively. The measured dissolved oxygen concentration and saturation values exhibit significant seasonal fluctuation. These results show that the outlet measurements change more than the inlet measurements.

The metal concentrations varied between 1.24 and 2.561 g/L (mean 1.799 g/L) for Cr1, 0.586 and 1.094 g/L (mean 0.790 g/L) for Cu1, 0.0058 and 0.0121 g/L (mean 0.0087 g/L) for Cd1, and 0.397 and 0.521 g/L (mean 0.470 g/L) for Pb1 in the inlet water. Accordingly, the outlet water metal concentrations were measured
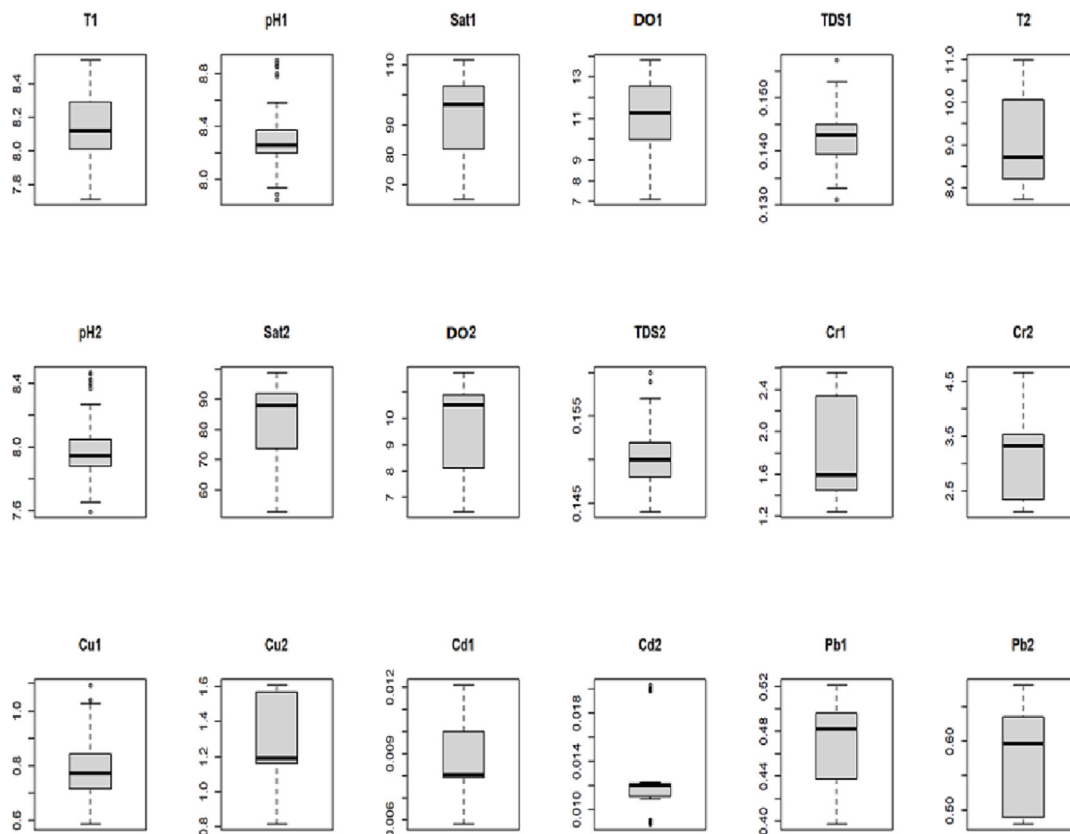
Fig. 1. Statistical scatterplot of all the features in the data set obtained from the fish farm (T1: Inlet water temperature (°C), pH1: Potential of hydrogen value of inlet water, Sat1: Inlet water saturation (%), DO1: Dissolved O_2 concentration (mg/L) of inlet water, TDS1: Total dissolved solids (g/L) in inlet water, T2: Outlet water temperature (°C), pH2: Potential of hydrogen value in outlet water, Sat2: Outlet water saturation (%), DO2: Dissolved O_2 concentration (mg/L) of outlet water, TDS2: Total dissolved solids (g/L) in outlet water, Cr1: Inlet water Chromium concentration (μg/L), Cr2: Chromium concentration (μg/L) in outlet water, Cu1: Inlet water Copper concentration (μg/L), Cu2: Copper concentration (μg/L) in outlet water, Cd1: Inlet water Cadmium concentration (μg/L), Cd2: Cadmium concentration (μg/L) in outlet water, Pb1: Inlet water Lead concentration (μg/L), Pb2: Lead concentration (μg/L) in outlet water).

in the range of 2.109–4.666 μg/L (mean 3.212 μg/L) for Cr2, 0.814–1.607 μg/L (mean 1.255 μg/L) for Cu2, 0.0088–0.0203 μg/L (mean 0.0127 μg/L) for Cd2 and 0.479–0.681 μg/L (mean 0.5788 μg/L) for Pb2 in the outlet water.

The measured characteristics (for dissolved oxygen, pH, and temperature) and metal availability (for Cr, Cu, Cd and Pb) for water environment were below the permissible limits set by official regulations (TR, 2006) for trout aquaculture.

3.1. Importance of features and feature selection algorithm

The graph of the feature importance ranking was obtained according to the XGBoost algorithm (Fig. 2). The saturation values of water (Sat1 and Sat2) had 100% and 82.5% scores, respectively, as the first two most important features. The Pb2 is in the third place with the importance score of 52%. T2 is in the fourth place in the feature importance ranking with a score of 26%. Three features with a score < 20% were DO2 14.24%, Cd1 6.5% and TDS2 4.98%. Apart from these features, Cr1, Pb1, Cu2, DO1, Cu1, Month, TDS1, Cd2, Cr2, pH1, T1 and pH2 features were non-significant in model building since they had 0% score values.

3.2. Hyper parameter optimization results for ML techniques

The Caret library creates logistic and linear regression models of neural networks with the multinom method. When constructing the LR model with this method, the weight decay (wd) hyper parameter is used and this parameter needs to be optimized

(Venables and Ripley, 2002). According to the results obtained according to the accuracy and kappa metrics (Table 4) for the wd value from zero to one, the most appropriate wd number value is 0.7.

According to Meyer (2022), the RBF-kernel is used when performing the classification task with SVM because it has fewer parameters and generally performs better. We performed repeated k-fold cross validation on training data to assess the quality of training results. The results shown in Table 5 were obtained in the hyper parameter optimization process using 18 support vectors with iterative cross validation. According to the results obtained with Accuracy and Kappa metrics, Sigma: 0.16542906187424 and Cost: 0.5 values were optimal hyper parameter values for SVM algorithm.

The results obtained after the hyper parameter optimization process for NB are shown in Table 6. One of tuning parameter, fL, was held constant at 0. Tuning parameter "adjust" was held constant at a value of 1. Accuracy was used to select the optimal model using the largest value. The final values used for the model were fL = 0, usekernel = TRUE and adjust = 1.

3.3. Model evaluation with metrics for ML techniques

For LR algorithm, first training success and then testing success evaluation were carried out using various metrics. Table 7 shows the status of training success according to the Accuracy and Kappa metrics of the models obtained by ML techniques.
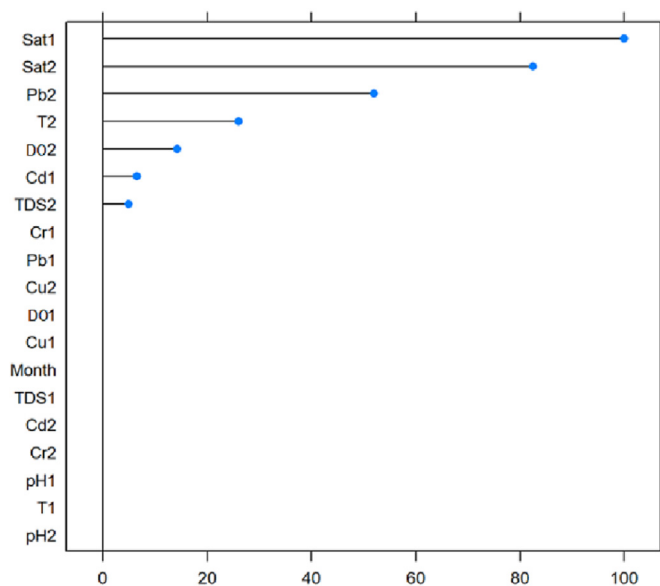
Variable Importance via XGBoost Method

Fig. 2. Features importance ranking by XGBoost algorithm (T1: Inlet water temperature (℃), pH1: Potential of hydrogen value of inlet water, Sat1: Inlet water saturation (%), DO1: Dissolved O2 concentration (mg/L) of inlet water, TDS1: Total dissolved solids (g/L) in inlet water, T2: Outlet water temperature (℃), pH2: Potential of hydrogen value in outlet water, Sat2: Outlet water saturation (%), DO2: Dissolved O2 concentration (mg/L) of outlet water, TDS2: Total dissolved solids (g/L) in outlet water, Cr1: Inlet water Chromium concentration (µg/L), Cr2: Chromium concentration (µg/L) in outlet water, Cu1: Inlet water Copper concentration (µg/L), Cu2: Copper concentration (µg/L) in outlet water, Cd1: Inlet water Cadmium concentration (µg/L), Cd2: Cadmium concentration (µg/L) in outlet water, Pb1: Inlet water Lead concentration (µg/L), Pb2: Lead concentration (µg/L) in outlet water).

Table 4
Tune table of weight decay for Logistic Regression algorithm.

decay	Accuracy	Kappa
0.001	0.907333	0.698119
0.01	0.912056	0.706601
0.1	0.946667	0.799249
0.2	0.951167	0.812513
0.4	0.963889	0.849391
0.5	0.968611	0.865217
0.6	0.968611	0.865217
0.7	0.970833	0.873044
0.8	0.968333	0.853044
0.9	0.968333	0.853044

When the Table 5 is evaluated, it is seen that the model created by SVM algorithm with training data is more successful than LR and NB algorithms based on the Accuracy and Kappa values.

The LR algorithm aims to obtain the optimum values of the β values, which are the coefficients of the selected features. The general equation of the resulting model can be seen in Eq. (18).

$$\textbf{Model} = \beta_0 + \beta_1 Sat1 + \beta_2 Sat2 + \beta_3 Pb2 + \beta_4 T2 + \beta_5 DO2 + \beta_6 Cd1 + \beta_7 TDS2 \tag{18}$$

After the β coefficients of the features were calculated with the LR algorithm, the model was obtained with residual deviance: 8.4801 and AIC: 24.4801 as shown in Eq. (19).

$$\textbf{Model} = 0.11 - 0.02 Sat1 - 0.169 Sat2 + 0.208 Pb2 + 1.3 T2 - 0.046 DO2 + 0.006 Cd1 + 0.022 TDS2 \tag{19}$$

The LR resulted in the best performance metrics for testing data set (accuracy: 0.867, Kappa: 0.659, sensitivity: 0.909, specificity: 0.75, precision: 0.909, prevalence: 0.733, F1: 0.909, p-value: 0.000194) after feature selection and hyper parameter optimization were applied. The confusion matrix and metrics are shown in Fig. 3.

As we can see from the results above, the p-value (0.000194) is smaller than the threshold value of 5%. This enables us to safely reject the null hypothesis and accept the alternate hypothesis. In other words, Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 have an impact on bacteria occurrence which allows us to conclude that Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 are a good predictor for bacteria occurrence in a trout farm.

The SVM resulted in the best performance metrics for testing dataset (accuracy: 0.933, Kappa: 0.815, sensitivity: 1, specificity: 0.75, precision: 0.917, prevalence: 0.733, F1: 0.909, p-value: 0.0006157) after feature selection and hyper parameter optimization were applied. The confusion matrix and metrics are shown in Fig. 3.

As we can see from the results above, the p-value is 0.0006157 which is smaller than the threshold value of 5%. This enables us to safely reject the null hypothesis and accept the alternate hypothesis. In other words, Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 have an impact on bacteria occurrence which allows us to conclude that Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 are a good predictor for bacteria occurrence in a trout farm.

The NB resulted in the best performance metrics for testing data set (accuracy: 0.867, Kappa: 0.706, sensitivity: 0.818, specificity: 1, precision: 1, prevalence: 0.733, F1: 0.9, p-value: 0.009539) after feature selection and hyper parameter optimization were applied. The confusion matrix and metrics are shown in Fig. 3.

Table 5
Tune table of sigma and cost hyper parameters with repeated cross validation for Support Vector Machine algorithm.

	Sigma	Cost	Accuracy	Kappa	Accuracy SD	Kappa SD
1	0.1654	0.2500	0.9673	0.8664	0.0545	0.2447
2	0.1654	0.5000	0.9775	0.9096	0.0455	0.2005
3	0.1654	1.0000	0.9775	0.9096	0.0455	0.2005
4	0.1654	2.0000	0.9775	0.9096	0.0455	0.2005
5	0.1654	4.0000	0.9775	0.9096	0.0455	0.2005

SD: Standard Deviation.

Table 6
Tune table of fL and adjust hyper parameters with repeated cross validation for Support Vector Machine algorithm.

	Usekernel	fL	adjust	Accuracy	Kappa	Accuracy SD	Kappa SD
1	FALSE	0	1	0.9319	0.7877	0.0769	0.2426
2	TRUE	0	1	0.9452	0.8201	0.0708	0.2314

fL: Factor for Laplace correction, SD: Standard Deviation. usekernel: if TRUE a kernel density estimate (density) is used for density estimation, If FALSE a normal density is estimated.

Table 7

Repeated cross-validation based models' comparison with accuracy and kappa metrics.

Accuracy						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LR	0.778	0.889	1	0.947	1	1
SVM	0.875	1.000	1	0.978	1	1
NB	0.667	0.889	1	0.945	1	1
Kappa						
	Min.	1st Qu.	Median	Mean	3rd Qu	Max.
LR	0	0.609	1	0.800	1	1
SVM	0	1.000	1	0.910	1	1
NB	0	0.609	1	0.820	1	1

Qu: Quartile, LR: Logistic Regression, SVM: Support Vector Machine, NB: Naïve Bayes, Max.: Maximum, Min: Minimum.

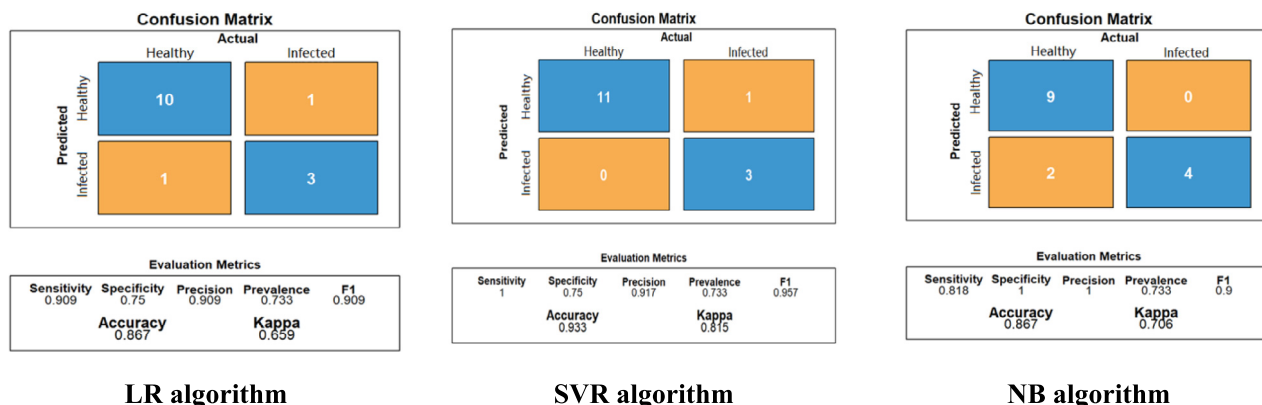**LR algorithm**   **SVR algorithm**   **NB algorithm**

Fig. 3. Confusion matrix and evaluation metrics for the best model of LR, SVR and NB algorithms (F1: F1 score metric).

As we can see from the results above, the p-value (0.009539) is smaller than the threshold value of 5%. This enables us to safely reject the null hypothesis and accept the alternate hypothesis. In other words, Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 have an impact on bacteria emergence which allows us to conclude that Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2 are a good predictor for disease outbreak caused by both bacteria (*Lactococcus garvieae* and *Vagococcus* sp.) in a trout farm.

4. Discussion

The XGBoost algorithm indicated that the disease emerged when the Pb2, T2, Cd1, and TDS2 values increased, and saturations (Sat1 and Sat2) and DO2 values decreased. The decrease in saturation and dissolved oxygen concentration seen in inlet and outlet water have the potential to cause the disease because of stress. Coldwater fish, such as trout, require 6 mg/L of dissolved oxygen to maintain its health (Abdel-Tawwab et al., 2019). The lowest measured value was close to this limit concentration (6.45 mg/L). Furthermore, earlier results of Matousek et al. (2017) indicated that fish survival rates decreased at 50%-60% oxygen saturation levels. Moreover, the toxic metals, such as Pb and Cd, alter immunoregulatory functions of fish. Thus, increased concentration of these metals in the aquatic environment may make the host more vulnerable to infections (Zelikoff, 1993). Although, temperature of the input water remains constant throughout the year, there are fluctuations in the temperature of fishponds. Even though the temperature change in fishponds varies by season. The XGBoost algorithm revealed that it was efficient in emergence of disease outbreak (Fig. 2). Fluctuations in water temperature have an effective role in the occurrence of the disease (Alborali, 2006). Besides, as in the present study, increased TDS concentrations indicate unhealthy water conditions for fish (Ahmed et al., 2019).

It is possible to increase the saturation and dissolved oxygen concentrations of water with appropriate aeration and oxygenation techniques. Thus, the stress caused by these two parameters on the cultured fish can be avoided. Similarly, filtration techniques such as millipore membrane (Kosemani et al., 2017) can reduce elemental load in water. However, such filtration approach must be employed efficiently and practically in a sustainable manner and the farm should utilize a recirculating aquaculture system. A similar scenario exists in case of water temperature stabilization.

After the feature importance ranking process was performed with the XGBoost algorithm, it was determined that the most important features in model building are Sat1, Sat2, Pb2, T2, DO2, Cd1 and TDS2. The model creation process will continue with these seven features in the next stages. We were unable to locate any study evaluating the joint effects of water quality characteristics and water metal load on the emergence of disease outbreaks with ML in the literature. Fish disease detection (Ahmed et al., 2022), weight estimation (Fernandes et al., 2020), fish counting (French et al., 2015), biomass detection (Abinaya et al., 2022) and sex discrimination (Barulin, 2019) have been reported in the earlier studies. Moreover, several studies (Ren et al., 2020; Huan et al., 2020; Cao et al., 2020) have successfully predicted the dissolved oxygen concentration in water with ML. Furthermore, the influence of water quality on aquatic livestock (growth, survival, and yield) has been investigated with ML (Rana et al., 2021).

The dataset was modeled with three popular ML techniques. These models were evaluated with test dataset and seven different metrics using repeated cross validation. All the models produced successful results and the most successful model was SVM (93.3% accuracy). The possibility of water quality characteristics revealing a disease outbreak was predicted by multinomial logistic regression analysis with 95.65% accuracy rate (Yilmaz et al., 2022). Similarly, fish diseases could be early detected via water quality index

with an average of 92% accuracy by using the gradient boosting model (Nayan et al., 2021). These results are compatible with the present study. In these studies, disease outbreak was previously predicted. With the assistance of this prediction, the emergence of the disease can be avoided by taking precautionary measures to fix the factors that may cause the disease. Thus, disease-related losses can be avoided.

Many reported studies focused on fish disease identification via ML techniques such as image processing (Ahmed et al., 2022; Mia et al., 2022). These kinds of studies are related to diseased fish detection or disease identification at early stages, but not to prediction of disease outbreak in advance. Contrary to the present study, the information from these investigations can be used to diagnose disease at an early stage and begin therapy.

5. Conclusion

Disease outbreaks in fish farms cause huge monetary losses. Changes in physico-chemical characteristics in the aquatic environment such as pH, dissolved oxygen concentration and stressors such as metals may alter fish susceptibility to diseases brought on by pathogens. The fish farming environment must be of high quality for healthy fish. To support fish welfare, an effective disease management can be carried out by monitoring the changes in the culture environment instantly. The techniques of ML, which is one of the sub-fields of artificial intelligence, can help to build effective aquaculture monitoring and management systems. The capability to monitor changes in the aquaculture environment is made possible by advances in computer technology. Technology allows for the monitoring of circumstances that cause significant losses in the industry as well as access to information that allows for early intervention. Furthermore, using ML to produce more accurate estimations of what affects fish production in fish farm will allow firms to make better decisions in advance. The ability of making an early intervention to improve product quality can be provided by smart production systems. In the aquaculture industry, more accurate risk management and sustainable production can be achieved with intelligent aquaculture production systems established by employing computer technology.

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.

Acknowledgements

Ethical statement

The procedures applied in this study were evaluated by the Akdeniz University Animal Experiments Local Ethics Committee and their ethical compliance was approved with the protocol number 2018.03.001. No substantial part of this study was published or considered for publication elsewhere.

References

Abdel-Tawwab, M., Monier, M.N., Hoseinifar, S.H., Faggio, C., 2019. Fish response to hypoxia stress: growth, physiological, and immunological biomarkers. Fish Physiol. Biochem. 45 (3), 997–1013. https://doi.org/10.1007/s10695-019-00614-9.

Abinaya, N.S., Susan, D., Sidharthan, R.K., 2022. Deep learning-based segmental analysis of fish for biomass estimation in an occulted environment. Comput. Electron. Agric. 197 (106985), 1–15. https://doi.org/10.1016/j.compag.2022.106985.

Ahmed, M.S., Aurpa, T.T., Azad, M.A.K., 2022. Fish disease detection using image based machine learning technique in aquaculture. J. King Saud Univ. Compt. Inf. Sci. 34, 5170–5182. https://doi.org/10.1016/j.jksuci.2021.05.003.

Ahmed, M., Rahaman, M. O., Rahman, M., Kashem, M. A., 2019, December. Analyzing the quality of water and predicting the suitability for fish farming based on IoT in the context of Bangladesh. In 2019 Int. Conf. on Sustain.Tech. for Indust. 4.0 (STI) 1-5. IEEE. https://doi.org/10.1109/STI47673.2019.9068050.

Ahsan, M.M., Luna, S.A., Zahed, S., 2022. Machine-learning-based disease diagnosis : a comprehensive review. Healthcare 10 (541), 1–30. https://doi.org/10.3390/healthcare10030541.

Alan, A., 2020. Evaluation of performance metrics and test techniques on various data sets in machine learning classification methods. Fırat University, pp. 1–82. Master's thesis.

Alborali, L., 2006. Climatic variations related to fish diseases and production. Vet. Res. Commun. 30 (Suppl. 1), 93–97. https://doi.org/10.1007/s11259-006-0019-7.

Alshari, H., Saleh, A.Y., Odabaş, A., 2021. Comparison of gradient boosting decision tree algorithms for CPU performance. Erciyes Univ. Jour. of Inst. of Scien. and Tech. 37 (1), 157–168.

Barulin, N.V., 2019. Using machine learning algorithms to analyse the scute structure and sex identification of sterlet Acipenser ruthenus (Acipenseridae). Aquac. Res. 50 (10), 2810–2825. https://doi.org/10.1111/are.14233.

Baydemir, M., 2021. The importance of statistics and choosing the right method in scientific researches. Bitlis Eren Univ. Soc. Sic. J. 10 (1), 29–34. https://doi.org/10.47130/bitlissos.919607.

Berrar, D., 2018. Bayes' theorem and naive Bayes classifier. In: Ranganathan, S., Nakai, K., Schonbach, C. (Eds.), Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics. Elsevier, pp. 403–412.

Binkhonain, M., Zhao, L., 2019. A review of machine learning algorithms for identification and classification of non-functional requirements. Expert Sys. Appl: X 2019 (100001), 1–23. https://doi.org/10.1016/j.eswax.2019.100001.

Bryan, G.W., Darracott, A., 1979. Bioaccumulation of marine pollutants. Sci. 286 (1015), 483–505. https://doi.org/10.1098/rstb.1979.0042. Philos. Trans. R. Soc. of Lond., B, Biol..

Cao, X., Liu, Y., Wang, J., Liu, C., Duan, Q., 2020. Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network. Aquc. Eng. 91 (102122), 1–10. https://doi.org/10.1016/j.aquaeng.2020.102122.

Cotes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1109/64.163674.

Diler, O., Altun, S., Adiloglu, A.K., Kubilay, A., Isıklı, B., 2002. First occurrence of Streptococcosis affecting farmed rainbow trout (Oncorhynchus mykiss) in Turkey. Bull. Eur. Assoc. Fish Pathol. 22 (1), 21–26.

Fernandes, A. F., Turra, E. M., de Alvarenga, E . R., Passafaro, T. L., Lopes, F. B., Alves, G. F. O., Singh, V., Rosa, G. J. M., 2020. Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia. Comput. Electron. Agric. 170(105274), 1-10. https://doi.org/10.1016/j.compag.2020.105274.

French, G., Fisher, M., Mackiewicz, M., Needle, C., 2015. Convolutional neural networks for counting fish in fisheries surveillance video. Proc. Mach. Vis. Anim. Their Behav. (MVAB), BMVA Press 7 (1), 1–10.

Friedman, J., 2001. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 29 (5), 1189–1232.

Hansen, G.H., Olafsen, J.A., 1999. Bacterial interactions in early life stages of marine cold water fish. Microb. Ecol. 38 (1), 1–26. https://doi.org/10.1007/s002489900158.

Hossin, M., Sulaiman, M., 2015. A Review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manag. Process 5 (2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201.

Huan, J., Li, H., Li, M., Chen, B., 2020. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of Chang Zhou fishery demonstration base. China. Comput. Electron. Agric. 175 (105530), 1–9. https://doi.org/10.1016/j.compag.2020.105530.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An introduction to statistical learning with applications in R (2nd ed.). Springer Science+Business Media, LLC. https://doi.org/10.1007/978-1-0716-1418-1.

Kosemani, S., Adewole, H., Olaleye, V., Azeez, Y., Saanu, C., Kosemani, E., 2017. Assessment of effect of fish feeding practices on the water quality of some fish ponds in Ekiti State Fish Farm, Ado Ekiti. Nigeria. Int. J. Fish. Aquat. Stu. 5, 357–364.

Kumar, G., Banerjee, R., Kr Singh, D., Choubey, N., Arnaw, 2020. Mathematics for Machine Learning. J. Math. Sci. Comput. Math. 1 (2), 229–238. https://doi.org/10.15864/jmscm.1208.

Kusuda, R., Kawai, K., 1998. Bacterial diseases of cultured marine fish in Japan. Fish Pathol. 33 (4), 221–227.

Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors. 18 (8), 1–29. https://doi.org/10.3390/s18082674.

Matousek, J., Prokesova, M., Novikava, K., Sebesta, R., Zuskova, E., Stejskal, V., 2017. The effect of water oxygen saturation on growth and haematological profile of juvenile peled Coregonus peled (Gmelin). Aquac. Res. 48 (10), 5411–5417. https://doi.org/10.1111/are.13356.

McHugh, M.L., 2012. Interrater reliability: The kappa statistic. Biochem. Med. 22 (3), 276–282. https://doi.org/10.11613/bm.2012.031.

Meyer, D., 2022. Support Vector Machines. RPubs by Rstudio, p. 8. Retrived from: Wien: R-Project.org.

Mia, M.J., Mahmud, R.B., Sadad, M.S., Al Asad, H., Hossain, R., 2022. An in-depth automated approach for fish disease recognition. J. King Saud Univ. - Comp. Inf. Sci. 34 (9), 7174–7183. https://doi.org/10.1016/j.jksuci.2022.02.023.

Nayan, A.A., Saha, J., Mozumder, A.N., Mahmud, K.R., Al Azad, A.K., Kibria, M.G., 2021. A machine learning approach for early detection of fish diseases by analyzing water quality. Trends Sci. 18 (21), 1–11. https://doi.org/10.48048/tis.2021.351.

Nieboer, E., Richardson, D.H., 1980. The replacement of the nondescript term 'heavy metals' by a biologically and chemically significant classification of metal ions. Environ. Pollut. Ser. B, Chemic. Phys. 1 (1), 3–26.

Önder, H., Cebeci, Z., 2014. Variable selection in logistic regression. Çukurova Univ. J. of Fac. Agri. 17 (2), 105–114.

Powers, D.M.W., 2011. Evaluation: from precision, recall and f-measure to ROC, ınformedness, markedness and correlation. J. Mach. Learn Tech. 2 (1), 37–63. https://doi.org/10.48550/arXiv.2010.16061.

Prabhakaran, S., 2016. Performance analysis and companion functions for binary classification models. Retrieved from http://r-statistics.co/Information-Value-With-R.html#3.4. IV.

Rana, M., Rahman, A., Dabrowski, J., Arnold, S., McCulloch, J., Pais, B., 2021. Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds. Biosyst. Eng. 208, 164–175. https://doi.org/10.1016/j.biosystemseng.2021.05.017.

Ren, Q., Wang, X., Li, W., Wei, Y., An, D., 2020. Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network. Aquac. Eng. 90 (102085), 1–13. https://doi.org/10.1016/j.aquaeng.2020.102085.

Rish, I., 2001. An empirical study of the Naive Bayes classifier. IJCAI 2001 Works. Empiric. Meth. Artif. Intell., 3(22), 41-46. https://doi.org/10.1039/b104835j.

Ruiz-Zarzuela, I., de Bias, I., Gironés, O., Ghittino, C., MúAzquiz, J.L., 2005. Isolation of vagococcus salmoninarum in rainbow trout, oncorhynchus mykiss (walbaum), broodstocks: characterization of the pathogen. Vet. Res. Commun. 29 (7), 553–562.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M.A., Shambour, M.K.Y., Alsalibi, A.I., Gandomi, A.H., 2022. Machine learning in medical applications: A review of state-of-the-art methods. Comput. Biol. Med. 145 (105458), 1–21. https://doi.org/10.1016/j.compbiomed.2022.105458.

Taheri, S., Mammadov, M., 2013. Learning the naive bayes classifier with optimization models. Int. J. Appl. Math. Comp. Sci. 23 (4), 787–795. https://doi.org/10.2478/amcs-2013-0059.

Tchounwou, P.B., Yedjou, C.G., Patlolla, A.K., Sutton, D.J., 2012. Heavy metal toxicity and the environment. Molecular, Clinic. Env. Toxicol., 133–164 https://doi.org/10.1007/978-3-7643-8340-4_6.

TR (Turkish Regulation), 2006. Application principles regarding aquaculture regulations (Republic of Turkey Ministry of agriculture and forestry, circular 2006/1). Retrived from: https://www.tarim orman.gov.tr/Belge ler/Mevzu at/ Genel geler /2006_1gene lge.pdf.

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S. Springer, New York.

Verschuere, L., Rombaut, G., Sorgeloos, P., Verstraete, W., 2000. Probiotic bacteria as biological control agents in aquaculture. Microbiol. Mol. Biol. Rev. 64 (4), 655–671. https://doi.org/10.1128/MMBR.64.4.655-671.2000.

Vupa Çilengiroğlu, Ö., Yavuz, A., 2020. Comparison of predictive performance of logistic regression and CART methods for life satisfaction data. Eur. J. Sci. and Technol. 18, 719–727. https://doi.org/10.31590/ejosat.691215.

Williams, R.J.P., 1996. Aluminium and biological systems: an introduction. Coord. Chem. Rev. 149, 1–9. https://doi.org/10.1016/S0010-8545(96)90004-1.

Winton, J.R., 2001. Fish health management. Fish Hatch. Man., 559–640

Yilmaz, M., Çakir, M., Oral, O., Oral, M.A., Arslan, T., 2022. Using machine learning technique for disease outbreak prediction in rainbow trout (Oncorhynchus mykiss) farms. Aquac. Res. 53, 6721–6732. https://doi.org/10.1111/are.16140.

Yin, M., Vaughan, J.W., Wallach, H., 2019. Understanding the effect of accuracy on trust in machine learning models. Conf. Hum. Fact. Comp. Syst. 1–12. https://doi.org/10.1145/3290605.3300509.

Zelikoff, J.T., 1993. Metal pollution-induced immunomodulation in fish. Ann. Rev. Fish Dis. 3, 305–325. https://doi.org/10.1016/0959-8030(93)90041-9.