Food Analytical Methods (2024) 17:1134–1148
https://doi.org/10.1007/s12161-024-02635-4

RESEARCH

# Hyperparameter Optimized Rapid Prediction of Sea Bass Shelf Life with Machine Learning

**Remzi Gürfidan[1] · İsmail Yüksel Genç[2] · Hamit Armağan[3] · Recep Çolak[4]**

Received: 21 March 2024 / Accepted: 10 May 2024 / Published online: 18 May 2024


## Abstract

The article focuses on the importance of sea bass, which is preferred by consumers in Turkey and worldwide. However, seafood can deteriorate rapidly under unfavorable conditions during storage due to their nutrient content, water content, and weakness in connective tissues. Temperature changes, inappropriate processing methods during transportation, and temperature changes during storage in markets are reported to cause losses in seafood quality. The deterioration of seafood, especially in seafood stored under inappropriate conditions because of temperature, causes changes contrary to consumer preferences because of the rapid growth of microorganisms, especially odor changes in seafood. This study examines the models related to the discipline of predictive microbiology, which are stated to provide an accurate shelf life prediction of the rate of microbiological spoilage and emphasize the importance of mathematical predictions of these models for seafood. Furthermore, the paper observes that machine learning algorithms such as Random Forest, Decision Tree, k-Nearest Neighbors, AdaBoost, Gradient Tree Boosting, Random Forest, Decision Tree, k-Nearest Neighbors, AdaBoost, and Gradient Tree Boosting have been used to predict the shelf life of seafood products. Finally, how to augment the limited data in a laboratory study to evaluate the shelf life of sea bass stored at different temperatures, how to prove the consistency of the augmented data with the original data, and how to optimize successful machine learning methods for robust problem-solving processes between different engineering fields are explained in detail. The results show that the optimized Extra Tree algorithm is the most successful for *Pseudomonas* quantity estimation with an $R^2$ metric value of 0.9940 and TVC quantity estimation with an $R^2$ metric value of 0.9910, while the other algorithms are less successful than this algorithm. These results show that machine learning methods can be a rapid, powerful, and effective tool for shelf life prediction of sea bass. Additionally, it should be emphasized that the number of input parameters (temperature, number of the bacteria) are of utmost significant for augmentation of the data for development and application of the machine learning algorithms.

**Keywords** Seabass · Shelf life · Extra Trees · Machine learning · *Pseudomonas* · TVC

## Introduction

Seafood has become one of the most important and growing sectors in meeting the protein needs in the world. In the world today, the nutritional content that people need plays an important role in increasing the demand for seafood (Anagnostopoulos et al. 2022; Odeyemi et al. 2018). Sea bass is one of the most preferred aquaculture products by consumers in Turkey as in many countries around the world. Nevertheless, Turkey has a very important position in sea bass farming and sea bass production in 2021 and 2022 was reported as 155,151 and 156,602 tons, respectively (Çöteli̇, 2023). Seafood products are among the perishable foods, and they can start to deteriorate very quickly during storage under unsuitable conditions due to their nutrient content, water content,

✉ Remzi Gürfidan
remzigurfidan@isparta.edu.tr

1 Computer Programming, Yalvaç Vocational School of Technical Sciences, Isparta University of Applied Sciences, Isparta, Turkey

2 Department of Fishing and Processing Technology, Eğirdir Faculty of Fisheries, Isparta University of Applied Sciences, Isparta, Turkey

3 Süleyman Demirel University, Rectorate, Turkey

4 Computer Programming, Vocational School of Distance Education, Isparta University of Applied Sciences, Isparta, Turkey

and weakness in connective tissues. Moreover, quality losses in seafood products are highly dependent on temperature fluctuations during transportation, inappropriate processing methods, and temperature changes during storage in markets (Alparslan et al. 2012; Turan and Kocatepe 2013; Masniyom et al. 2002).

Spoilage of seafood results from the rapid proliferation of microorganisms in seafood stored under inappropriate conditions, especially with the effect of temperature, causing changes contrary to consumer preferences, especially odor in seafood. It has been reported that spoilage microorganisms in seafood generally belong to *Pseudomonas* sp, Enterbacteriaceae, *Vibrio* sp, Lactobacillaceae, and Bacillaceae groups (Gram and Dalgaard, 2002; Gram 2009). The rate of microbiological spoilage in seafood can provide accurate shelf life estimates. Predictive microbiology is a discipline that studies models that can predict shelf life according to the growth rates of microorganisms. Shelf life prediction of seafood products according to mathematical models is of great importance in terms of time and food safety (Messens et al. 2018). Many mathematical models have been developed for shelf life prediction in seafood. Koutsoumanis (2001) studied microbial growth in sea bream under storage conditions between 0 and 15 °C. *Pseudomonas* sp. was considered the specific spoilage microorganism and bacterial growth was analyzed according to the Balehradek type model. According to the results of the study, the maximum specific growth rate (μmax) and the minimum temperature of the lag phase (Tmin) were reported as −11.8 and −12.8 °C, respectively. In addition, the bias and accuracy factors of the developed model were reported to be between 0.91 and 1.17. In another study with sea bream, samples naturally contaminated with *Pseudomonas* sp, *Shewanella putrefaciens*, Enterobacteriaceae, lactic acid bacteria, and yeasts were examined. Sea bream was stored between 0 and 15 °C and the growth of microorganisms was analyzed. The researchers reported that *Pseudomonas* sp. is a good indicator of spoilage. According to the results of the research, it was reported that the developed model provided realistic and accurate results under the specified storage conditions (Koutsoumanis and Nychas 2000).

On the other hand, machine learning algorithms that can predict the quality and shelf life of seafood products are also used with predictive microbiology. When the studies were reviewed, it was seen that algorithms such as Random Forest, Decision Tree, k-Nearest Neighbors, AdaBoost, and Gradient Tree Boosting can predict the quality of seafood. Wijaya et al. (2023) aimed to determine the quality of seafood by using electronic noise and machine learning algorithms with hyperparameter optimization. The researchers examined different machine learning algorithms used in their study and reported that the k-Nearest Neighbors algorithm achieved the highest accuracy factor in predicting the quality of seafood. Moreover, they reported that the RMSE and $R^2$ values in the regression model were 0.03 and 0.995, respectively. Another study aimed to develop an electronic nose to determine the quality and shelf life of cultured Pacific white shrimp. The researchers performed pH, Total Volatile Basic Nitrogen (TVBN), Fourier Transform Infrared spectra (FTIR), and texture, microbiological, and sensory analyses to determine the quality and shelf life of Pacific white shrimp samples stored with and without ice. The researchers used Principal Component Analysis (PCA), Decision Tree, Random Forest, k-Nearest Neighbor (KNN), and Soft-max Regression in pattern recognition algorithms and reported that Soft-max Regression produced 96 and 95% decision accuracy for samples stored with and without ice, respectively (Srinivasan et al. 2020). Food storage conditions, drying, and preservation under different conditions are very important for human health. Researchers have used machine learning algorithms and optimization methods to predict the shelf life and spoilage times of food (Kaveh et al. 2023). This can sometimes translate into missions such as determining product quality and determining the accuracy of the operations performed on the product. When this is the goal, algorithms such as multiple regression, artificial neural network, and CNN are frequently preferred (Tito Anand et al. 2022). In the perspective of literature, the aim of this study is to evaluate the shelf life prediction of whole sea bass stored at different temperatures by using different machine learning algorithms and to investigate the practical use of the tested models.

Predicting the shelf life of seafood products is crucial for maintaining their quality and safety throughout the supply chain. Over the years, various mathematical models and machine learning techniques have been developed to forecast the shelf life of different marine fish species under varying storage conditions. This research aims to contribute to this field by focusing on the prediction of sea bass shelf life using advanced machine learning algorithms. Leveraging the insights from existing studies, such as Koutsoumanis et al. (2002) and Tran et al. (2019) who developed predictive models for fish shelf life and recent advancements like the work of Cui et al. (2024) who utilized machine learning for shelf life prediction across multiple marine fish species, this study aims to optimize hyperparameters to enhance the accuracy and speed of prediction specifically tailored for sea bass. Machine learning algorithms offer a promising approach for rapidly predicting the shelf life of seafood products like sea bass due to their ability to handle complex data patterns and nonlinear relationships. Drawing on the methodologies and findings from studies like García et al. (2022) and Yin et al. (2022), which focused on mathematical modeling and quality changes in fish storage, and An et al. (2023), who investigated the impact of packaging and storage conditions on fish quality, this research seeks to develop a robust

predictive model that can accommodate various factors influencing sea bass shelf life. By integrating insights from both traditional food science approaches and cutting-edge machine learning techniques, this study aims to provide a comprehensive framework for rapid and accurate shelf life prediction, facilitating better decision-making in seafood industry operations and ensuring the delivery of high-quality products to consumers.

## Laboratory Analyses

### Materials

At the total 34 whole sea bass with an average weight of 288.14 ± 45.32 g were used in this study. The samples were obtained from a commercial institution and were brought to the laboratory immediately after they were obtained in a drained styrofoam box with ice preservation. After the samples were brought to the laboratory, they were stored under aerobic conditions at 4, 10, and 19 °C for 12, 10, and 5 days, respectively. Two sea bass samples were taken periodically on days 0, 1, 2, 5, 8, and 12 for 4 °C; on days 0, 1, 3, 6, 8, and 10 for 10 °C; and finally 0, 1, 2, 3, and 5 for 19 °C and the samples were evaluated microbiologically. *Pseudomonas* sp. and Total Viable Count (TVC) values were analyzed to determine the microbiological quality (Fig. 1).

### Microbiological Analysis

Ten grams of skinned muscle samples was taken on each sampling day from whole sea bass stored aerobically at different temperatures and 10-fold dilutions were prepared accordingly. *Pseudomonas* sp. were counted on Cephaloridin-Fucidin-Cetrimide (CFC, Merck) agar with CFC supplement while TVC was determined on Plate Count Agar (CFC, Merck). For the incubation of the microorganisms, *Pseudomonas* sp. was counted after 48 h at 25 °C and TVC was determined at 30 °C after 48 h of incubation (Chuesiang et al. 2020; Poli et al. 2006).

## Data Augmentation Technique with Synthetic Data Generation Method

Initially, our dataset consisted of only 15 rows, which is not big enough for machine learning algorithms and insufficient for them to perform sound learning. Therefore, we used curve fitting to increase the size of your dataset. Time, temperature, and log data were augmented in our dataset. These are often the basic types of data that machine learning models need, and diversifying and augmenting this data allows your model to learn better. Our data augmentation process started with a manual configuration of the data types. In this step, we determined which data types to augment and how to augment them. This configuration step forms the basis of the data

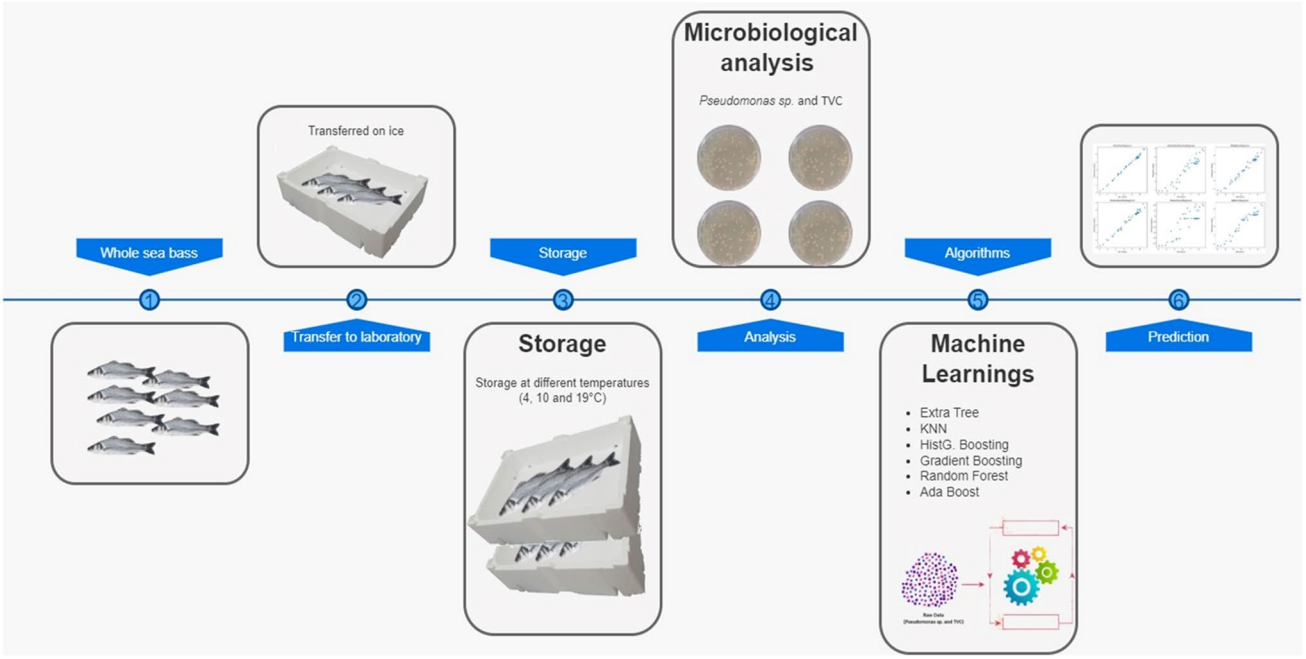



**Fig. 1** Schematic view of microbiological analyses and application of machine learning algorithms

augmentation process and ensures that the correct data types are identified. Next, the data types identified before starting the data augmentation were validated by matching them with the new data types to be created. This step is important to ensure that the correct data type mapping is done and that the data augmentation process is started correctly. In your original dataset, time values were located at specific hours, such as 4, 12, and 19. The missing integer values between these hours were filled by completing the series within your dataset. This makes the time data more seamless and continuous. The temperature values have been increased to show each hour between 4 and 19. This allows us to model the relationship between the available temperature data, resulting in data points covering a wider range of temperatures. Similarly, time values were increased to show each temperature value between 0 and 288 h. This allowed us to model the relationship between time and temperature in more detail. The pseudo code of the data augmentation method is shown in Algorithm 1 table.

**Algorithm 1** Data Augmentation Pseudo Code

```
function Curve_Fitting (time, temp, log)
    if (time, temp, log).length < 2 then
        return 0
    curve_params ← Initialize_Curve_Params()
    fitted_curve ← Fit_Curve(time, temp, log, curve_params)
    return fitted_curve
function Initialize_Curve_Parameters()
    curve_params ← Default_Curve_Params()
    return curve_params
function Fit_Curve(time, temp, log, curve_params)
    fitted_curve ← Curve_Fitting_Algorithm(time, temp,
log, curve_params)
    return fitted_curve
```

Consequently, the reason for increasing the data is to replicate the learning material so that machine learning models can produce better results (Maharana et al. 2022). The basic principle of data augmentation is to ensure that the distribution in the original dataset is highly correlated with the distribution in the synthetic dataset and that statistical distributions such as standard deviation and variance are similar.

## Trained Machine Learning Model and Regression

In this study, a total of six different machine learning prediction algorithms were trained. The success metrics and test graphs obtained because of these trainings were interpreted and the most successful algorithm from the experiments was determined as the Extra Tree Regressor model. The inputs of the machine learning model are time and temperature values. Based on these values, TVC and *Pseudomonas* values are estimated. The results obtained provide powerful data for the prediction of the shelf life of sea bass.

### Extra Tree Regressor Algorithm

The Extra Trees (Extremely Randomized Trees) algorithm is a machine learning method specifically used to solve classification and regression problems. Extra Trees is a method based on decision trees. It uses many trees like the Random Forest algorithm as a working logic. In addition, unlike Random Forest, Extra Trees takes more randomness into account when constructing trees (Breiman 2001; Geurts et al. 2006). $G_{\text{th}}$ denotes the prediction tree. Here, $\theta$ denotes a uniform independent distribution vector that is assigned before the growth of the tree. All trees are combined and averaged into a tree ensemble of $G(x)$, which is generated using the Breiman 2001 equation (Eq. 1) (Hammid et al. 2018).

**Table 1** Hyperparameters tried to be optimized and values tested

| | |
|---|---|
| **Parameters and their values tested for hyperparameter optimization** | n_estimators': [50, 100, 200]<br>'max_depth': [None, 10, 20, 30]<br>'min_samples_split': [2, 5, 10]<br>'min_samples_leaf': [1, 2, 4] |
| **Hypermeter values** | max_depth': 20<br>'min_samples_leaf': 2<br>'min_samples_split': 2<br>'n_estimators': 100 |

$$G\left(x,\theta_1,\cdots,\theta_2\right)=\frac{1}{2}\sum_{r=1}^{R}G\left(x,\theta_r\right) \tag{1}$$

GridSearchCV is a hyperparameter tuning method available in the scikit-learn library. It is used to experiment with various combinations of hyperparameters used to improve the performance of a model. By specifying a given hyperparameter space (parameter combinations), it evaluates the performance of the model for different combinations in that space and selects the hyperparameters that perform best. GridSearchCV tries to select the best hyperparameters by cross-validating over the specified hyperparameter combinations. In this study, the GridSearchCV method was applied to the most successful Extra Tree algorithm and the best values of the selected hyperparameters were determined. These best parameter values were then used to train the model. The tested and found hyperparameter values are shown in Table 1.

## k-Nearest Neighbors Regressor Algorithm

k-Nearest Neighbors (KNN) is an effective machine learning method that is preferred as a classification or regression solver. The algorithm uses the classes or values of the nearest neighboring points to classify or predict a new data point. The basic principle of KNN proceeds by recognizing that data points with similar characteristics tend to have the same class or a similar value. Considering $x$ and $y$ as axis values, after calculating the distance, the input $x$ is considered the class value with the highest probability. This is calculated by Eq. 2.

$$P(y=j|X=x)=\frac{1}{k}\sum_{i\epsilon A}I\left(y^{(i)}=j\right) \tag{2}$$

## HistGradient Boosting Regressor Algorithm

HistGradient Boosting Regressor is a machine learning algorithm available in the scikit-learn library. It is a type of Gradient Boosting algorithm and is specifically designed to work effectively on large datasets. HistGradient Boosting Regressor uses a histogram-based method to process data faster. This makes it more effective, especially on large datasets. In the equation, $y(x)$ represents the predicted target variable; $K$ represents the total number of trees. $f_{k(x)}$ is the prediction of each tree.

$$y(x)=\sum_{k=1}^{K}f_{k(x)} \tag{3}$$

Each tree focuses on correcting the errors of the previous trees and is a type of regression tree. These trees work by splitting the data and making a regression estimate for each region.

## Gradient Boosting Regressor Algorithm

Gradient Boosting is a machine learning algorithm used as a solver in classification and regression processes. This algorithm aims to create a strong learner by combining weak learners together. Gradient Boosting aims to combine weak predictors (usually decision tree type models) to create a strong prediction model. The basic principle of how this algorithm works is to correct the erroneous learning of the previous weak estimator by adding new estimators. This process affects the calculation of the weights, while the new values are determined by the loss function. Equation 4 is used for the overall model calculation.

$$\gamma_m=arg_\gamma min\sum_{i}^{n}L\left(y_i,F_{m-1}\left(x_i\right)+\gamma\right) \tag{4}$$

Here, $i=1$-n belongs to $r_{ij}$, where $j$ represents the leaf, $y$ is the observed value, and γ is the predicted value.

## Random Forest Regressor Algorithm

Random Forest is a machine learning algorithm that is widely used especially in classification and regression problems. Random Forest can create a more powerful and generalizable model by combining multiple decision trees. When decision trees are configured for regression models, the average of the decision trees is the prediction value. Random Forest uses randomization to minimize the risk of overfitting. Random feature selections and random generation of

data subsets make the model more diverse and generalizable. Mean square error value for Random Forest is calculated as in Eq. 5.

$$RF_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \quad (5)$$

where $N$ is the number of data points, $f_i$ is the value returned by the model, and $y_i$ is the actual value for data point $i$.

### AdaBoost Regressor Algorithm

AdaBoost (Adaptive Boosting) is an ensemble learning algorithm for building strong models. AdaBoost aims to build a stronger model by combining weak models together. The AdaBoost algorithm is an algorithm that works on weights and each weak classifier is assigned a weight. Once a classifier is trained on the weighted training set, the weights of the misclassified examples are increased and the next classifier is trained on this updated weighted data set. This process continues until a desired number of iterations or specific learning objective is reached. Equation 6 is used for the overall model calculation.

$$\epsilon_t = \frac{\sum_{i=1}^{N} w_{i.1(h_t(x_i)\neq y_i)}}{\sum_{i=1}^{N1} w_i} \quad (6)$$

The error rate is calculated by $\varepsilon_t$, that is, it shows how well the $t$th classifier is able to correct the errors made on the weighted training data set.

### Evaluation of the Models

Error metrics used to evaluate the success of machine learning algorithms are used to measure how well the model performs. These metrics help to assess how well a model's predictions match the true values and the generalization ability of the model.

Mean absolute error (MAE) is a metric that shows how close the predicted values are to the true values. This metric is calculated by Eq. 7 (Hammid et al. 2018; Mishra et al. 2017; AlOmar et al. 2020).

$$MAE = \frac{1}{n}\sum_{r=1}^{n}\left|P_d^{r,m} - P_d^{r,c}\right| \quad (7)$$

Root means square error (RMSE) was chosen to compare the prediction errors of different trained models. The closer the RMSE value is to 0, the better the predictive ability of the model in terms of its absolute deviation. The RMSE value is calculated by Eq. 8 (Hammid et al. 2018; Mishra et al. 2017; AlOmar et al. 2020; Willmott and Matsuura 2005).

$$RMSE = \sqrt{\frac{1}{n}\sum_{r=1}^{n}(P_d^{r,m} - P_d^{r,c})^2} \quad (8)$$

The coefficient of determination ($R^2$) is used to estimate model efficiency and is calculated by Eq. 9 (Hammid et al. 2018).

$$R^2 = 1 - \frac{\sum_{r=1}^{n}(P_d^{r,m} - P_d^{r,c})^2}{\sum_{r=1}^{n}(P_d^{r,m} - P_d^{-r,m})^2} \quad (9)$$

MSE either assesses the quality of an estimator. The MSE metric is calculated by Eq. 10.

$$MSE = \frac{1}{n}\sum_{r=1}^{n}(P_i - P_i')^2 \quad (10)$$

## Result and Discussions

### Microbiological Results

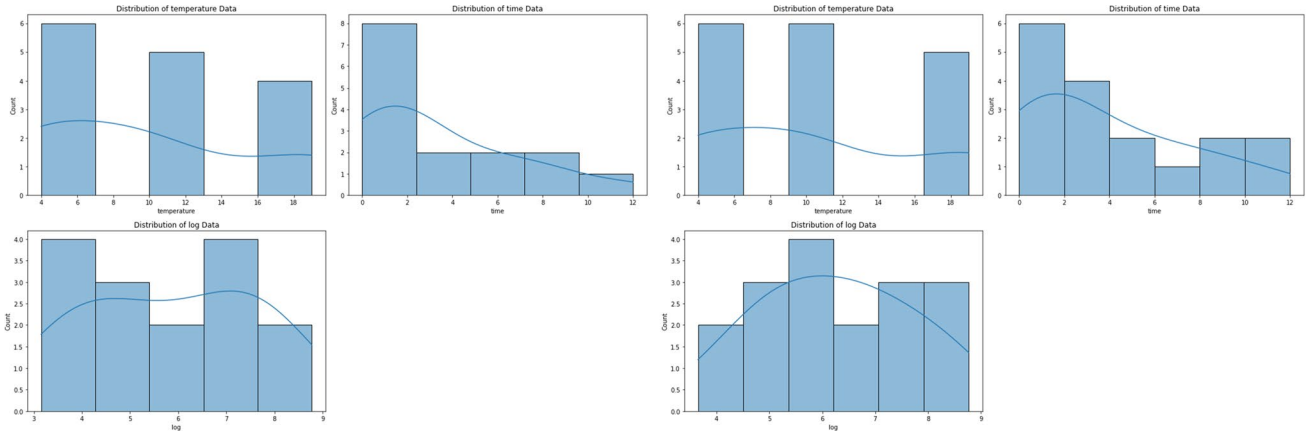
Microbiological changes in whole sea bass stored at different temperatures are shown in Table 2.

According to the results of the research, the initial number of the bacteria varied between 3.15 and 4.03 log cfu/g in whole sea bass stored at different temperatures. The highest initial value was observed at 10 °C. *Pseudomonas* sp. numbers increased with storage time and the highest value was reported as 8.65 log cfu/g. In a study, quality changes of whole and filleted sea bass samples stored on ice were investigated during 16 days of storage. The researchers reported that the initial value of *Pseudomonas* sp. was higher (1.4–2.8 log cfu/g) in whole seabass than in filleted seabass. Furthermore, they reported that the value of 7 log cfu/g was exceeded on the 8th day of storage in filleted seabass, whereas the number of *Pseudomonas* sp. was 6.4 log cfu/g on the 16th day of storage in whole seabass (Taliadourou et al. 2003). Compared to this study in which different machine learning algorithms were evaluated for shelf life prediction, it was observed that the initial microflora of whole seabass was similar but the maximum number of the

**Table 2** Microbiological quality changes of whole seabass stored at different temperatures

| Bacteria | Storage temperature (°C) | $N_{min} - N_{max}$ (log cfu/g)* |
|---|---|---|
| *Pseudomonas* sp. | 4 | 3.15–7.65 |
| | 10 | 4.03–8.65 |
| | 19 | 3.72–8.10 |
| Total Viable Count | 4 | 4.19–8.20 |
| | 10 | 4.95–7.23 |
| | 19 | 3.73–8.27 |

* $N_{min}$ and $N_{max}$ representing the minimum and maximum number of the bacteria




a) Pseudomonast b) TVC

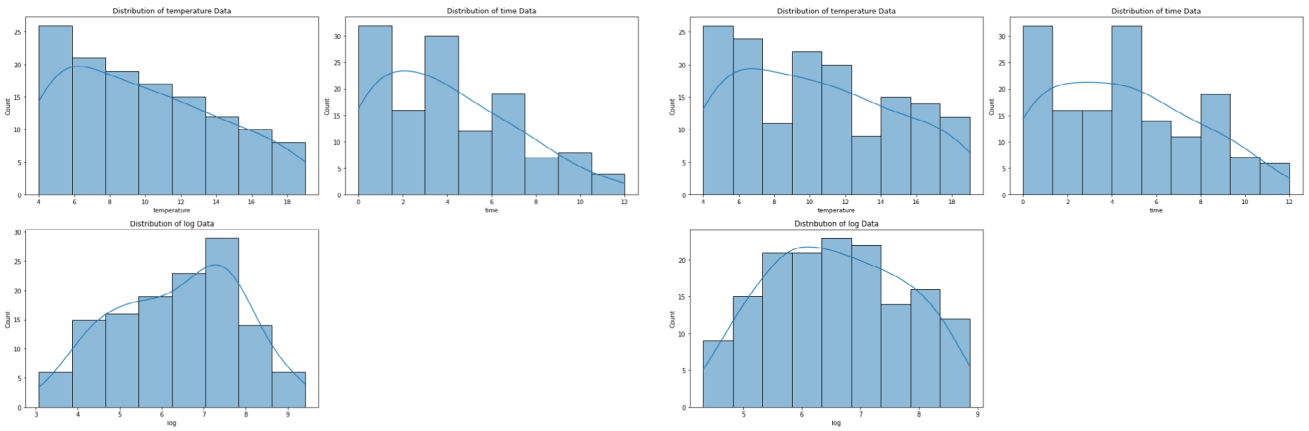
**Fig. 2** Original data set scatter plot

bacteria was higher and it was considered that this was due to the differences in storage temperatures. In another study, Ntzimani et al. (2023) investigated the effect of slurry ice on transportation and storage quality in seabass. The researchers reported an initial *Pseudomonas* sp. value of $2.0 \pm 0.1$ log cfu/g in seabass, lower than the values obtained in this study where different algorithms were tested.

## Reliability and Validity Analysis of the New Data Set

In cases where the original data set is insufficient for machine learning algorithms to learn, researchers try to increase the data set. An important criterion to be considered in this process is that the new data to be produced in the process of increasing the number of data have qualities as similar as possible to the original data set. In order to monitor this situation, researchers perform some tests at the end of the process and compare the measurements. Figure 2 shows the scatter plot of the original data set, where a part of the data set is *Pseudomonas* sp. and b part of the data set is TVC bacteria.

Figure 3 shows the scatter plot of the new data set after the data augmentation was applied. When analyzed figuratively, it is seen that the trend line on the temperature, time, and log graphs are quite close to each other in both. The small differences observed in the column graphs will be explained by the numerical values measured in the similarity metrics. Data augmentation has been reported by researchers to be a method that plays a role in predicting data at points that are not tested and in making more accurate and reliable classification (Georgouli et al. 2018). In a study, Prema and Visumathi (2022) developed a non-destruction method for shrimp using hybrid CNN and SVM with Generative Adversarial Network (GAN)




a) *Pseudomonas* b)TVC

**Fig. 3** Augmented dataset scatterplot

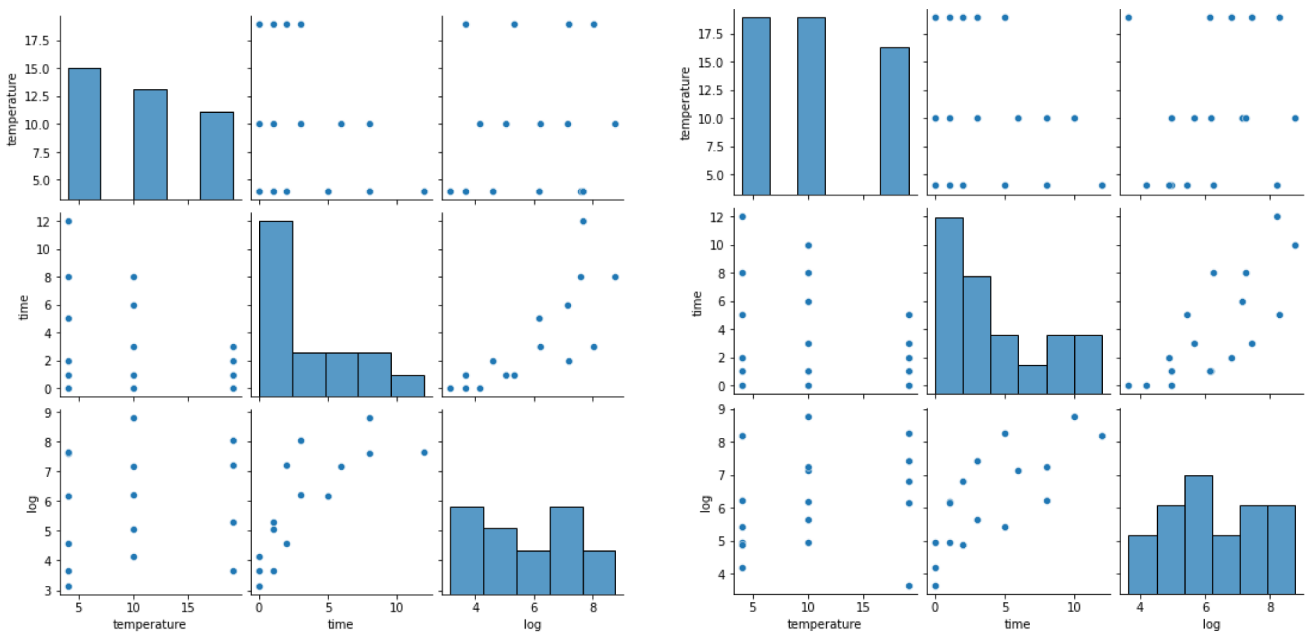



**Fig. 4** Original dataset correlation matrix

augmentation. The researchers reported that CNN and CVM methods had an accuracy of 98.1% in the new data set obtained by augmentation of the original data with GAN. In the present study, as seen in Fig. 3, it is seen that the new data set obtained has become more accurate in the machine learning algorithms used in the study. In this context, it is observed that data augmentation improves the performance of the algorithms used.

The temperature distribution in bands 4, 12, and 19 shown in Fig. 2 is filled in Fig. 3 for all values between 4

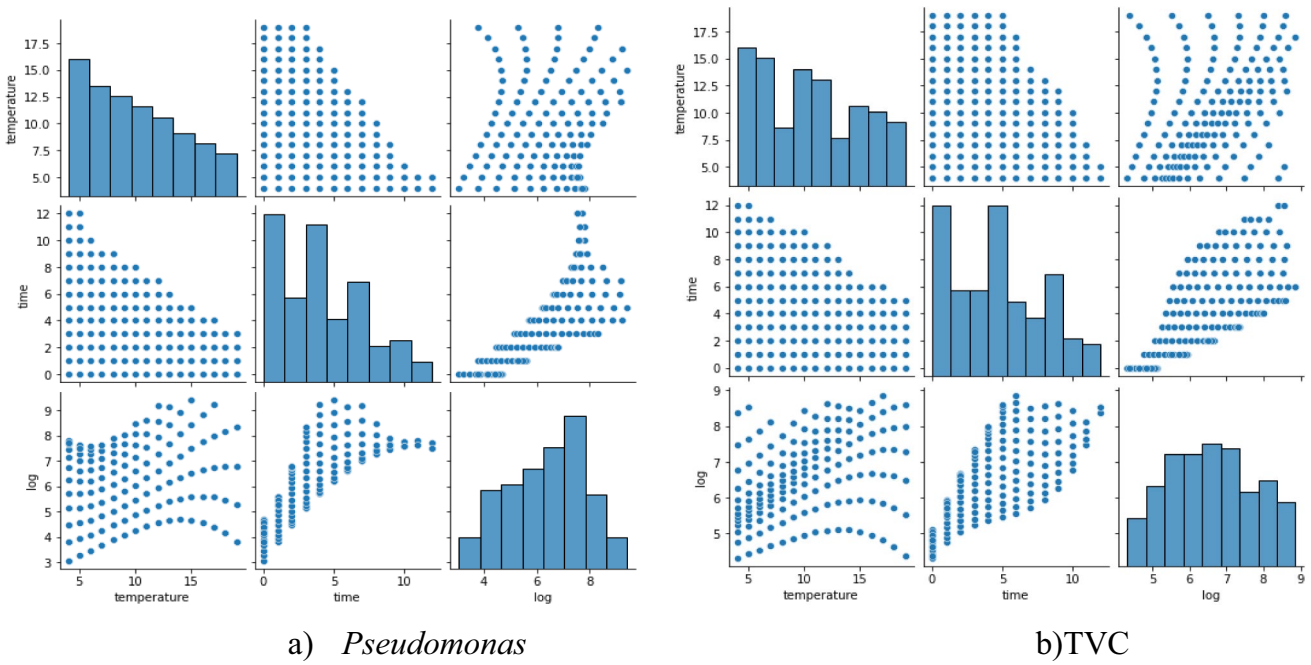



**Fig. 5** Augmented dataset correlation matrix

and 19. In addition, the slope line in Fig. 2 and the slope line in Fig. 3 are similar to each other. This method was repeated for time and log variables. The scatter and slope plots for Pseudomonas and TVC are similar between Figs. 2 and 3.

Figure 4 shows the correlation graph of the original dataset, showing the effects of the parameters on each other. The sets of graphs, aggregated into column and dot plots, show the row-column cross interactions of temperature, time, and log data.

Figure 5 shows the correlation graph showing the effects of the parameters of the new augmented dataset on each other. The sets of graphs, aggregated into column and dot plots, show the row-column cross interactions of temperature, time, and log data.

When the formation of the part graphs in Fig. 5 is carefully examined, it is seen that the graphs are formed by staying within the frames of the part graphs drawn in Fig. 4. This examination shows us the similarity of the effects of the parameters in the old and new data set on each other. As a general conclusion, the similarity of the new data set with the original data set is proved by the correlation matrix graphs. Table 3 shows the results of the measurements performed on the original and new augmented data set values in terms of number of data, mean value, standard deviation, minimum, first quartile, half, last quartile, and maximum metrics.

As shown in Table 3, the closeness of the numerical values in the findings indicates the consistency of the generated data with the original data. As for the metrics for *Pseudomonas*sp., the max deviation between augmented and original data was found in the number of the bacteria in which the original count of bacteria was 8.77 while the augmented data were found to be 9.41. However, the differences between the original and augmented data were not significant as the growth of the bacteria reached its stationary phase.

Table 4 shows the success and error metric values of the machine learning models after training. Thanks to these values, the most successful machine learning model for this dataset can be determined. The $R^2$ metric takes values

**Table 3** Mathematical inferences from the original dataset vs. augmented dataset

| | Metrics | Original temperature | Augmented temperature | Original time | Augmented time | Original log | Augmented log |
|---|---|---|---|---|---|---|---|
| *Pseudomonas* | Count | 15 | 128 | 15 | 128 | 15 | 128 |
| | Mean | 10.00000 | 9.890625 | 3.466667 | 4.000000 | 5.886192 | 6.380892 |
| | Std | 6.21059 | 4.401564 | 3.602909 | 3.053332 | 1.808930 | 1.498693 |
| | Min | 4.00000 | 4.000000 | 0.000000 | 0.000000 | 3.150515 | 3.072504 |
| | 25% | 4.00000 | 6.000000 | 1.000000 | 1.750000 | 4.361790 | 5.239686 |
| | 50% | 10.00000 | 9.000000 | 2.000000 | 3.500000 | 6.150515 | 6.613867 |
| | 75% | 14.50000 | 13.000000 | 5.500000 | 6.000000 | 7.396083 | 7.592421 |
| | Max | 19.00000 | 19.000000 | 12.000000 | 12.000000 | 8.778151 | 9.415678 |
| TVC | Count | 17 | 153 | 17 | 153 | 17 | 153 |
| | Mean | 10.529412 | 10.372549 | 3.941176 | 4.575163 | 6.243559 | 6.583693 |
| | Std | 6.206022 | 4.461653 | 3.732804 | 3.205039 | 1.479044 | 1.149553 |
| | Min | 4.000000 | 4.000000 | 0.000000 | 0.000000 | 3.650515 | 4.320041 |
| | 25% | 4.000000 | 7.000000 | 1.000000 | 2.000000 | 4.951545 | 5.706496 |
| | 50% | 10.000000 | 10.000000 | 2.000000 | 4.000000 | 6.190106 | 6.585960 |
| | 75% | 19.000000 | 14.000000 | 6.000000 | 7.000000 | 7.238561 | 7.498807 |
| | Max | 19.000000 | 19.000000 | 12.000000 | 12.000000 | 8.772034 | 8.864692 |

**Table 4** Test results and error metric values of machine learning algorithms

| Bacteria type | *Pseudomonas* | | | | TVC | | | |
|---|---|---|---|---|---|---|---|---|
| Algorithm and Metrics | $R^2$ | RMSE | MSE | MAE | $R^2$ | RMSE | MSE | MAE |
| **Extra Tree** | 0.9940 | 0.1069 | 0.0114 | 0.0783 | 0.991 | 0.085 | 0.0073 | 0.055 |
| **KNN** | 0,9733 | 0.2338 | 0.0542 | 0.1722 | 0.960 | 0.190 | 0.0362 | 0.123 |
| **HistG. Boosting** | 0.9058 | 0.4377 | 0.1916 | 0.3604 | 0.827 | 0.395 | 0.156 | 0.266 |
| **Gradient Boosting** | 0.9817 | 0.1926 | 0.0371 | 0.1428 | 0.979 | 0.135 | 0.018 | 0.112 |
| **Random Forest** | 0.7395 | 0.7282 | 0.5302 | 0.6413 | 0.504 | 0.671 | 0.450 | 0.567 |
| **AdaBoost** | 0.9311 | 0.3744 | 0.1402 | 0.3275 | 0.894 | 0.310 | 0.096 | 0.260 |

between 0 and 1. A value of 1 indicates that the model can explain all variations, while 0 indicates that the independent variable of the model cannot explain the dependent variable significantly. For this reason, the value sought should be as close to 1 as possible. When the values in Table 4 are interpreted, it is seen that the Extra Trees algorithm has the closest $R^2$ value to 1. This result shows that the algorithm has a very successful prediction capability. In addition, the low MAE, MSE, and RMSE values obtained are important numerical evidence supporting the success of the model. The RMSE metric shows that the model predicts closer to the actual values as it approaches 0. As can be seen from Table 4, Extra Tree algorithm has the smallest $R^2$ value. As the MSE metric approaches 0 which is 0.0114 for *Pseudomonas*sp. and 0.0073 for TVC, the model is considered to perform better compared to other tested algorithms.

In a study, it was aimed to develop an electronic nose for cultured Pacific white shrimp. The researchers examined the odor formation during storage in Pacific white shrimp stored at 2 °C with and without ice. Pattern recognition algorithms based on multivariate analysis, such as Principal Component Analysis (PCA), Decision Tree, Random Forest, K-Nearest Neighbor (KNN) and Soft-max Regression were used in the study. In addition, it was reported that the Soft-max Regression algorithm showed about 96% and 95% accuracy for Pacific white shrimp stored on ice and without ice, respectively (Srinivasan et al. 2020). In another study, Gowda et al. (2023) investigated the determination of freshness in edible seafood using IoT and machine learning techniques. The researchers tested machine learning algorithms such as Dense Net Algorithm and Efficient Net Algorithm while determining the freshness of the seafood products they used. According to the results of the research, Efficient Net and Dense Net Algorithm reported accuracy rates of 0.085 and 0.075, respectively. Compared to the studies in the literature, it was observed that machine learning algorithms produced successful results in determining the quality, shelf life prediction, and freshness classification of seafood products, although there were methodological differences in this study. In this study, it was observed that Extra Tree was the most successful algorithm in determining the numbers of *Pseudomonas* sp. and TVC with respect to storage time. However, it was further reported that the algorithms with the highest correlation for *Pseudomonas* sp. were Gradient Boosting, KNN, AdaBoost, HistG. Boosting, and Random Forest and for TVC, Extra Tree, Gradient Boosting, KNN, AdaBoost, HistG. Boosting, and Random Forest, respectively.

Figure 6 shows the graphs of the success of the different machine learning models trained. The blue colored dots on the graphs are the intersection point of the values on the $x$ and $y$ axes in the graph. In this type of graph, it is expected that the predicted value of the model and the actual value will lie linearly as a line on the $x$ and $y$ axis. The scattered appearance of the points on the axis indicates that the prediction success of the model worsens. As can be seen from the graphs, the most successful results were obtained with the Extra Tree algorithm. Yu et al. (2019) used deep learning and hyperspectral imaging to predict TVC in peeled Pacific shrimp. The researchers extracted hyperspectral features from near-infrared (NIR) hyperspectral imaging (HSI) using stacked auto-encoders (SAE) and developed a model to predict the TVC of peeled Pacific shrimp stored at 4 °C using a fully connected neural network (FNN). According to the results of the research, they reported the accuracy of predicting the TVC numbers of peeled shrimp during storage at 4 °C with deep learning algorithms as $R^2p = 0.927$. Compared to this study with sea bass, it is seen that the algorithms used in TVC are Extra Tree, KNN, and Gradient Boosting with higher $R^2$. For *Pseudomonas* sp., the algorithms with the highest $R^2$ were reported to be Extra Tree, KNN, Gradient Boosting, and AdaBoost. (Table 4; Fig. 6). Table 4 shows the test results and error metrics obtained from the machine learning algorithms' prediction of TVC and *Pseudmonas* sp. values. Figure 6 visualizes the consistency of the prediction of TVC and *Pseudmonas* sp. values with the actual values.

Figure 7 shows the prediction error distributions of the different machine learning models trained. The blue bar graphs on the graphs show how often errors occur at which values. When interpreting these graphs, the most successful situation is to identify the graph with the lowest possible number of error distributions. It is then important to be able to reduce the error frequencies. Thanks to this graph, enrichments can be made on the dataset by identifying data groups that have values that will enable the model to learn better or correct mislearning within the dataset in which the model is trained. As can be seen from this graph, the machine learning algorithm with the lowest error distribution and error frequency is the Extra Tree algorithm. In a study where e-nose data were used to classify the freshness of seafood products, the researchers used TVC growth as microbiological and e-nose data as sensory in three different seafood products such as sole fillets, red mullet fillets, and cuttlefish. The researchers used KNN and partial least square discriminant to classify the freshness of different seafood products used in the study. According to the results of the research, they reported that the KNN model showed 100% accuracy and stated that the model they developed can be used in seafood distribution centers (Grassi et al. 2022). In this study conducted in sea bass, it was determined that the $R^2$ value of the KNN algorithm was 0.97 for *Pseudomonas* sp. and 0.96 for TVC. However, it was revealed in this study that the prediction errors for *Pseudomonas* sp. and TVC was lowest in the Extra Tree algorithm. (Fig 6).

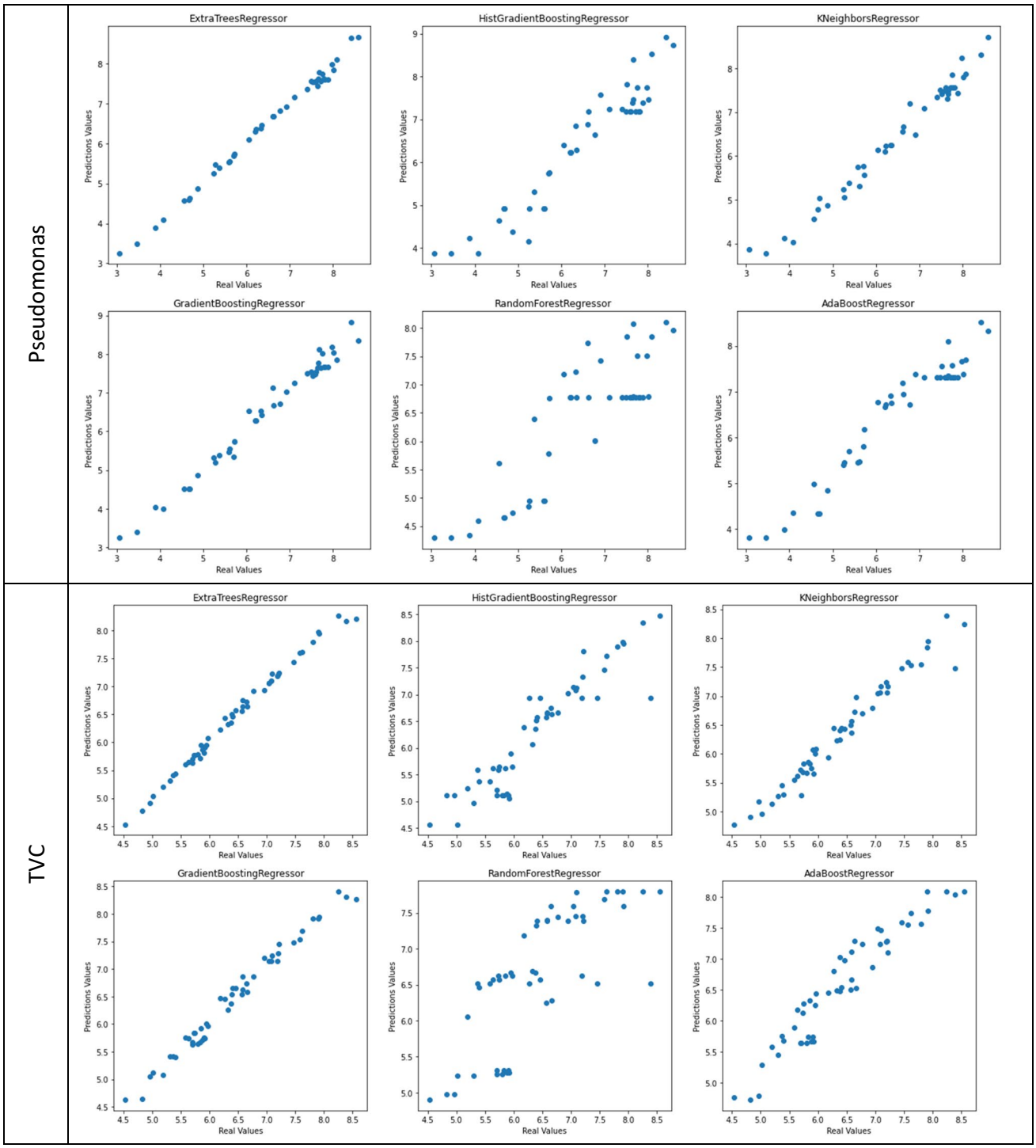



**Fig. 6** Success graphs of trained models

Figure 8 shows the graphs of the prediction success of the different machine learning models trained. The blue colored lines on the graphs show the actual values that should be predicted, and the orange lines show the predicted values of the models. The expected success image from these graphs is that the actual values and predicted values overlap. Wu et al. (2022) used convolutional neural network_ long short-term memory model to determine the freshness of salmon fillets at fluctuated temperatures. The researchers reported that CNN-LSTM model has better results than kinetic models such as logistic equation, Gompertz equation, and Arhenius equation. They also reported that the CNN-LSTM model had

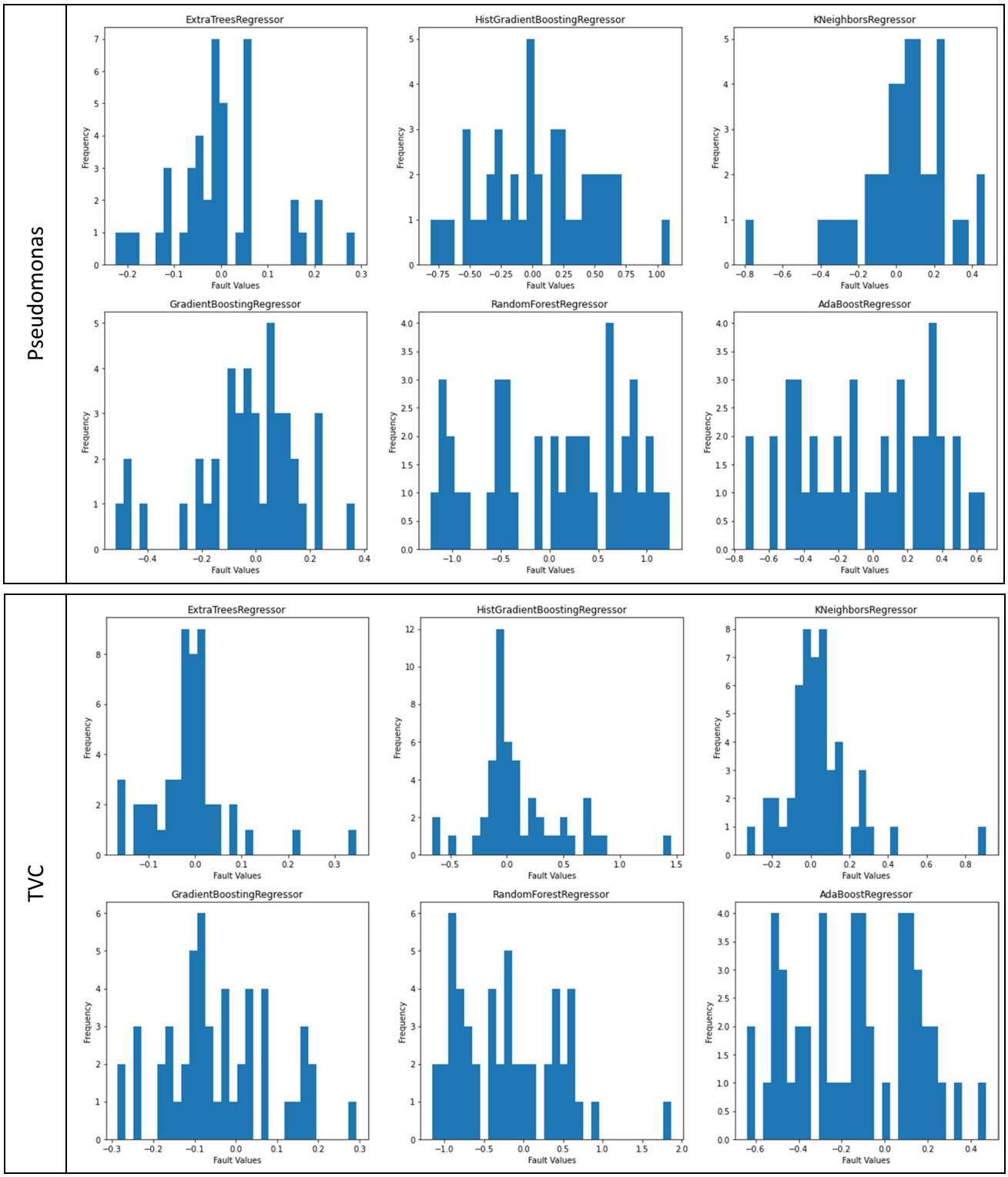



**Fig. 7** Prediction error scatter plots of the trained models

an $R^2$ value of 0.95 and an RMSE value of less than 0.2 at fluctuated temperatures in salmon fillets. In another study, freshness was estimated in frozen storage using e-nose, e-tongue, and colorimetric analysis in whole stored horse mackerel. The researchers used different machine learning algorithms such as artificial neural network, ANN; extreme gradient boosting, XGBoost; random forest regression, RFR; support vector regression, SVR and reported that ANN,

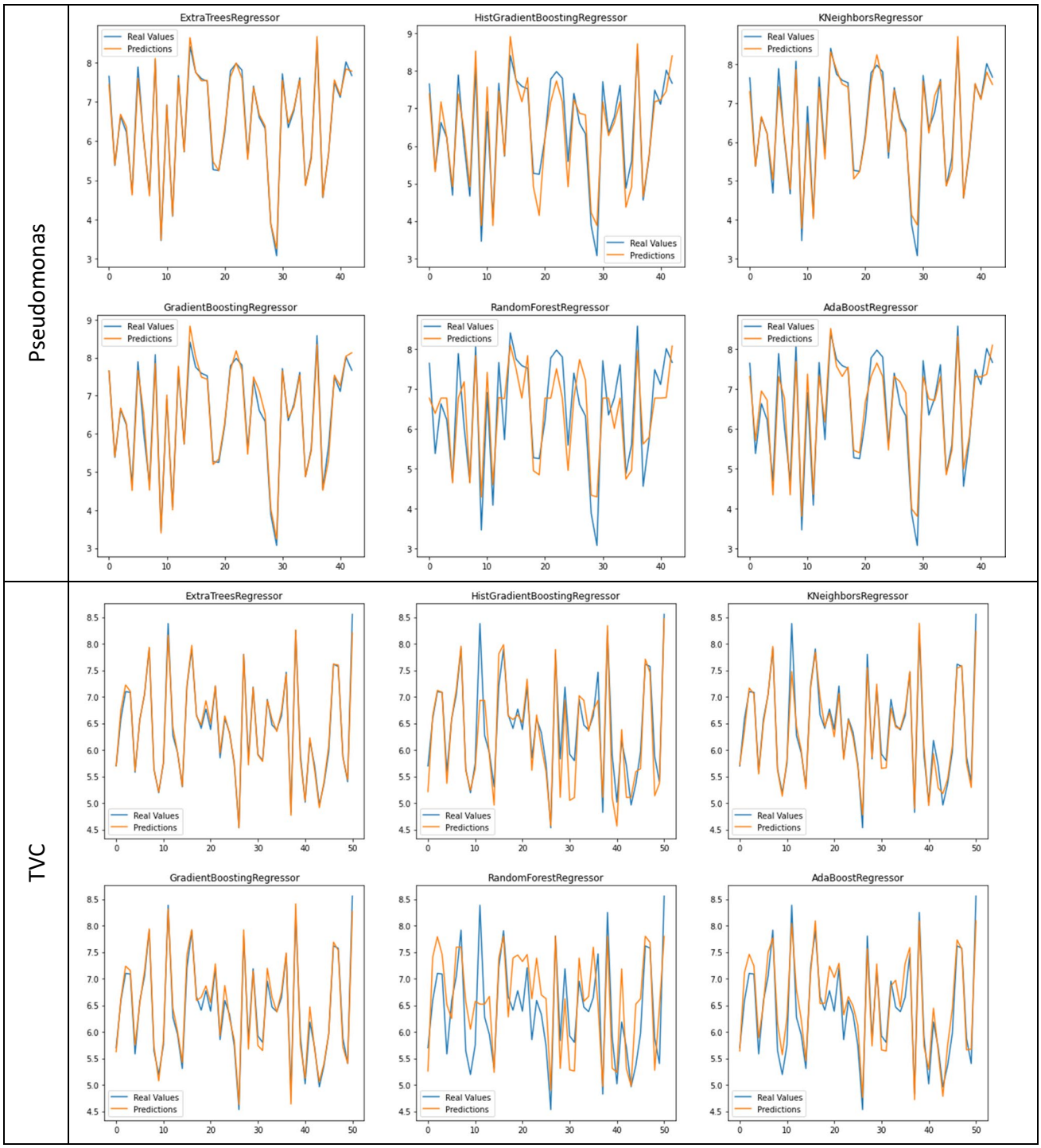



**Fig. 8** Prediction success value graphs of trained models

RFR, and XGBoost performed well in predicting biochemical indices (Li et al. 2023). In this study with sea bass, it was observed that the best performance belonged to the Extra Tree algorithm as shown in Fig. 8. Figure 8 visualizes the prediction success of TVC and *Pseudomonas* values during the repeated training process.

The values obtained in Fig. 8 are the results obtained from training on the augmented dataset. The success of the actual values shown in Fig. 8 in representing the experimental data is shown in Table 3, where the augmented data are close to the original values.

## Conclusions

In this study, *Pseudomonas* sp. and TVC were analyzed during storage in whole sea bass stored at different temperatures (4, 10, and 19 °C). Data augmentation was performed using laboratory data and a shelf life prediction model was developed using different machine learning algorithms such as Extra Tree, Gradient Boosting, KNN, AdaBoost, HistG. Boosting, and Random Forest. According to the model performance evaluations, Extra Tree was the best performing algorithm ($R^2_{Pseudomonas} = 0.9940$ and $RMSE_{Pseudomonas} = 0.1069$; $R^2_{TVC} = 0.991$ and $RMSE_{TVC} = 0.085$). During this study, different data augmentation methods were tried. Techniques such as logistic regression, CTGAN, and curve fitting are among the methods tried. Among these techniques, the most efficient result was found in the curve fitting technique. The process of faithfully generating synthetic data is a critical issue for machine learning algorithms to produce successful results. In this study, we would like to emphasize the importance of synthetic data generation methods due to the difficulty of obtaining the studied data and the limitations of the study. This study shows that machine learning algorithms can successfully predict the shelf life of seafood. However, it should be noted that machine learning algorithms require large amounts of data and need to be developed separately for each seafood product. It is also suggested that more input data during the development of the model will bring more precise results. In future studies, it is concluded that increasing the bacterial groups and adding chemical parameters to the model can increase the precision and accuracy of the model.

**Author Contribution** Remzi Gürfïdan: machine learning methods and fine-tuning process, writing discussion and results. İsmail Yüksel Genç: microbiological experiments, writing introduction and conclusion. Hamit Armağan: writing introduction and related works. Recep Çolak: writing introduction and related works.

**Data Availability** No datasets were generated or analyzed during the current study.

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

AlOmar MK, Hameed MM, AlSaadi MA (2020) Multi hours ahead prediction of surface ozone gas concentration: robust artificial intelligence approach. Atmos Pollut Res 11(9):1572–1587. https://doi.org/10.1016/j.apr.2020.06.024

Alparslan Y, Gürel Ç, Metin C, Hasanhocaoğlu H, Baygar T (2012) Determination of sensory and quality changes at treated sea bass (Dicentrarchus labrax) during cold-storage. J Food Process Technol 3(183):2

An Y, Liu N, Xiong J, Li P, Shen S, Qin X, Huang Q (2023) Quality changes and shelf-life prediction of pre-processed snakehead fish fillet seasoned by yeast extract: affected by packaging method and storage temperature. Food Chem Adv 3:100418

Anagnostopoulos DA, Parlapani FF, Boziaris IS (2022) The evolution of knowledge on seafood spoilage microbiota from the 20th to the 21st century: have we finished or just begun? Trends Food Sci Technol 120:236–247

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chuesiang P, Sanguandeekul R, Siripatrawan U (2020) Phase inversion temperature-fabricated cinnamon oil nanoemulsion as a natural preservative for prolonging shelf-life of chilled Asian seabass (Lates calcarifer) fillets. Lwt 125:109122

Çöteli FT (2023) Agrıcultural economy and polıcy development instıtute tepge, product report aquaculture products 2023, 9. https://arastirma.tarimorman.gov.tr/tepge/Menu/37/Urun-Raporlari. Accessed 15 May 2024

Cui F, Zheng S, Wang D, Ren L, Meng Y, Ma R, Li J (2024) Development of machine learning-based shelf-life prediction models for multiple marine fish species and construction of a real-time prediction platform. Food Chem, 139230

García MR, Ferez-Rubio JA, Vilas C (2022) Assessment and prediction of fish freshness using mathematical modelling: a review. Foods 11(15):2312

Georgouli K, Osorio MT, Del Rincon M, J., Koidis A, (2018) Data augmentation in food science: synthesising spectroscopic data of vegetable oils for performance enhancement. J Chemom 32(6):e3004

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

Gowda B, Alamelu JV, Varsha K, Shetty A, Manjappa N (2023) Identification and detection of freshness in Edible fishes using Iot and Machine Learning techniques. J Surv Fisheries Sci 10(3):335–342

Gram L (2009) Microbiological spoilage of fish and seafood products. Compendium Microbiol Spoilage Foods Beverages, 87–119

Gram L, Dalgaard P (2002) Fish spoilage bacteria— problems and solutions. Curr OpinBiotechnol 13:262–266

Grassi S, Benedetti S, Magnani L, Pianezzola A, Buratti S (2022) Seafood freshness: e-nose data for classification purposes. Food Control 138:108994

Hammid AT, Bin Sulaiman MH, Abdalla AN (2018) Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using artificial neural network. Alexandria Eng J 57(1):211–221. https://doi.org/10.1016/j.aej.2016.12.011

Kaveh M, Çetin N, Khalife E, Abbaspour-Gilandeh Y, Sabouri M, Sharifian F (2023) Machine learning approaches for estimating apricot drying characteristics in various advanced and conventional dryers. J Food Process Eng 46(12):e14475. https://doi.org/10.1111/jfpe.14475

Koutsoumanis K (2001) Predictive modeling of the shelf life of fish under nonisothermal conditions. Appl Environ Microbiol 67(4):1821–1829

Koutsoumanis K, Nychas GJE (2000) Application of a systematic experimental procedure to develop a microbial model for rapid fish shelf life predictions. Int J Food Microbiol 60(2–3):171–184

Koutsoumanis K, Giannakourou MC, Taoukis PS, Nychas GJE (2002) Application of shelf life decision system (SLDS) to marine cultured fish quality. Int J Food Microbiol 73(2–3):375–382

Li H, Wang Y, Zhang J, Li X, Wang J, Yi S, Li J (2023) Prediction of the freshness of horse mackerel (Trachurus japonicus) using E-nose, E-tongue, and colorimeter based on biochemical indexes analyzed during frozen storage of whole fish. Food Chem 402:134325

Maharana K, Mondal S, Nemade B (2022) A review: data pre-processing and data augmentation techniques. *Global Transitions Proceedings* 3(1):91–99

Masniyom P, Benjakul S, Visessanguan W (2002) Shelf-life extension of refrigerated seabass slices under modified atmosphere packaging. J Sci Food Agric 82(8):873–880

Messens W, Hempen M, Koutsoumanis K (2018) Use of predictive modelling in recent work of the panel on Biological hazards of the European Food Safety Authority. Microb Risk Anal 10:37–43

Mishra G, Sehgal D, Valadi JK (2017) Hypothesis quantitative structure activity relationship study of the anti-hepatitis peptides employing Random forests and extra-trees regressors. Open Access Volume 13(3):60–62

Ntzimani A, Angelakopoulos R, Semenoglou I, Dermesonlouoglou E, Tsironi T, Moutou K, Taoukis P (2023) Slurry ice as an alternative cooling medium for fish harvesting and transportation: study of the effect on seabass flesh quality and shelf life. Aquaculture Fisheries 8(4):385–392

Odeyemi OA, Burke CM, Bolch CC, Stanley R (2018) Seafood spoilage microbiota and associated volatile organic compounds at different storage temperatures and packaging conditions. Int J Food Microbiol 280:87–99

Poli BM, Messini A, Parisi G, Scappini F, Vigiani V, Giorgi G, Vincenzini M (2006) Sensory, physical, chemical and microbiological changes in European sea bass (Dicentrarchus labrax) fillets packed under modified atmosphere/air or prepared from whole fish stored in ice. Int J Food Sci Technol 41(4):444–454

Prema K, Visumathi J (2022) An improved non-destructive shrimp freshness detection method based on hybrid CNN and SVM with GAN augmentation. In: 2022 international conference on advances in computing, communication and applied informatics (ACCAI). IEEE, pp 1–7

Srinivasan P, Robinson J, Geevaretnam J, Rayappan JBB (2020) Development of electronic nose (Shrimp-Nose) for the determination of perishable quality and shelf-life of cultured Pacific white shrimp (*Litopenaeus Vannamei*). Sens Actuators B 317:128192

Taliadourou D, Papadopoulos V, Domvridou E, Savvaidis IN, Kontominas MG (2003) Microbiological, chemical and sensory changes of whole and filleted Mediterranean aquacultured sea bass (Dicentrarchus labrax) stored in ice. J Sci Food Agric 83(13):1373–1379

Tito Anand MA, Anandakumar S, Pare A, Chandrasekar V, Venkatachalapathy N (2022) Modeling of process parameters to predict the efficiency of shallots stem cutting machine using multiple regression and artificial neural network. Journal of Food Process Engineering 45(6):e13944. https://doi.org/10.1111/jfpe.13944

Tran GD, Ndraha N, Hsiao HI (2019) Development of predictive model for the remaining shelf-life of tilapia fillet under variable temperature conditions. J Fish Soc Taiwan 46(1):19–30

Turan H, Kocatepe D (2013) Different MAP conditions to improve the shelf life of sea bass. Food Sci Biotechnol 22:1589–1599

Wijaya DR, Syarwan NF, Nugraha MA, Ananda D, Fahrudin T, Handayani R (2023) Seafood Quality Detection using electronic nose and Machine Learning Algorithms with Hyperparameter optimization. IEEE Access

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30(1):79–82. https://doi.org/10.3354/cr030079

Wu T, Lu J, Zou J, Chen N, Yang L (2022) Accurate prediction of salmon freshness under temperature fluctuations using the convolutional neural network long short-term memory model. J Food Eng 334:111171

Yin C, Wang J, Qian J, Xiong K, Zhang M (2022) Quality changes of rainbow trout stored under different packaging conditions and mathematical modeling for predicting the shelf life. Food Packaging Shelf Life 32:100824

Yu X, Yu X, Wen S, Yang J, Wang J (2019) Using deep learning and hyperspectral imaging to predict total viable count (TVC) in peeled Pacific white shrimp. J Food Meas Charact 13:2082–2094