Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Improving Fish Weight Estimation with Quantile and Box-Cox Transforms: Comparative Machine Learning Models

Hatice ESEN[1], Havvanur TAŞDELEN[2], Sefa KÜÇÜK[3], Işıl KARABEY AKSAKALLI[4]

[1] Erzurum Technical University, Department of Electrical-Electronics Engineering, hatice.esen96@erzurum.edu.tr, Orcid: 0009-0008-9805-829X
[2] Erzurum Technical University, Department of Electrical-Electronics Engineering, havvanur.tasdelen20@erzurum.edu.tr, Orcid: 0009-0000-8222-2093
[3] Erzurum Technical University, Department of Electrical-Electronics Engineering, sefa.kucuk@erzurum.edu.tr, Orcid: 0000-0002-0279-3185
[4*]Erzurum Technical University, Department of Computer Engineering, isil.karabey@erzurum.edu.tr, Orcid: 0000-0002-4156-9098

ARTICLE INFO

Article history:

Received 11 April 2025
Received in revised form 22 August 2025
Accepted 26 August 2025
Available online 30 September 2025

Keywords:

Machine learning, Quantile transformation, Box-cox transformation, Fish weight estimation.

Doi: 10.24012/dumf.1674123

* Corresponding author

ABSTRACT

Fish weight estimation using machine learning ensures that fish are fed appropriately, reduces labor, prevents physical harm to the fish, and saves time. In this study, Quantile and Box-Cox transformations are applied to improve the accuracy of fish weight predictions. These transformations correct the asymmetric distribution of the data and enable machine learning algorithms to generalize more effectively and produce more accurate results. CatBoost, Random Forest, Polynomial Regression, and Support Vector Regression methods were evaluated for fish weight estimation both before and after applying the transformations. The experimental results show that both the Quantile and Box-Cox transformations effectively reduce model error rates, particularly by normalizing the dataset distribution. Notably, models without transformation exhibit significant improvements in error rates after transformation is applied. The lowest Mean Absolute Error (MAE) without transformation was obtained using the CatBoost model, yielding a value of 14.002. After applying the Quantile transformation, the MAE decreased to 0.0171, while the Box-Cox transformation resulted in an MAE of 0.3302. Although both transformations contribute to error reduction, the Quantile transformation has a more substantial impact on fish weight estimation. These findings underscore the importance of data transformations in the preprocessing stage and highlight that transformation techniques are as crucial as selecting the appropriate machine learning model.

Introduction

Accurately estimating fish weight is crucial for nutrition planning and effective aquaculture management, as it facilitates the efficient regulation of feeding processes. Traditionally, fish weight is measured using weighing scales. However, these methods not only increase labor requirements but also fail to save time and may cause physical harm to the fish. Consequently, there is a growing demand for more accurate, non-invasive, and sustainable estimation techniques. Nonetheless, estimating weight in underwater environments remains a highly challenging task due to factors such as the continuous movement of fish, fluctuating lighting conditions, and variable water quality.

A review of recent literature reveals that machine learning (ML) and artificial intelligence (AI) have been extensively explored in academic research across various domains. For example, Wai Lok Woo et al. conducted a study that estimated the weight of Tilapia in turbid water using a low-cost single-channel video camera combined with a Mask R-CNN detection method, and subsequently employed regression models (Linear, Random Forest, SVR) to achieve high predictive accuracy [30]. For instance, Kazemi et al. conducted a comprehensive evaluation of ML algorithms used to predict the mechanical properties of fiber-reinforced polymers (FRPs), assessing model performance in detail [16]. Similarly, von Bülow et al. reviewed recent advances in ML-based approaches for investigating sequence–structure–function relationships in disordered proteins, emphasizing their relevance to biophysical functions [17]. In another study, Liu et al. developed a quantitative structure–property relationship (QSPR) model using ML techniques to predict CO_2 solubility in aqueous amine solutions. The study demonstrated high predictive accuracy and validated model interpretability through SHAP analyses [18]. D'Orazio and Pham examined the effects of climate-related financial policies on decarbonization and the renewable energy transition across 87 countries from 2000 to 2023, using ML to account for contextual differences and policy effectiveness [19]. Furthermore, Tuerxun et al. proposed an ML framework that integrates spectral indices and geospatial data to accurately estimate leaf chlorophyll

content (LCC), achieving high accuracy with the GWLS-Support Vector Regression (SVR) model [20].

For example, in a recent study, Tianye Zhang and colleagues combined fish posture recognition using deep learning technology with biomass estimation, developing a stereo vision-based fast, accurate, and fully automated system for free-swimming fish, and demonstrated that the method provides an effective approach for real production and the investigation of fish growth mechanisms [33]. Recent studies have demonstrated that image processing and machine learning techniques can be effectively applied to estimate fish weight. For instance, Gutzmann et al. proposed a method for estimating the weight of large whitefish species in the lower Mackenzie River Basin using photographic image analysis [1]. Another recent study by SV Jansi Rani et al. aimed to estimate fish biomass in highly turbid water using a deep learning-based object detection and regression approach. This study presents a non-invasive and automated method [34]. In another study, Moseli Mots'oehli et al. developed FishNet, which achieved 89% classification accuracy and a 2.3 cm MAE in fish length estimation using a dataset of 1.2 million fish images from 163 species [32]. Similarly, Konovalov et al. conducted a study on automatic weight estimation by collecting both images and weight measurements of approximately 2,500 individuals of Lates calcarifer (Asian sea bass or barramundi) harvested from three different locations in Queensland, Australia [2]. Additionally, Islamadina et al. utilized digital camera images converted to grayscale for fish weight estimation, applying segmentation techniques to remove irrelevant objects from the images. During the feature extraction phase, length, width, and height were calculated using calibration values, and these features were used to estimate fish weight [3]. A similar approach was adopted by Suwannakhun and Daungmala in their study on pig weight estimation. They extracted features such as color, texture, center of gravity, axis lengths, eccentricity, and area from digital images [4]. Neural network analyses in both studies confirmed the effectiveness of non-contact estimation methods. Collectively, these findings suggest that image processing and AI-based techniques provide faster, more accurate, and more sustainable alternatives to traditional weight measurement methods.

Quantile Transformation (QT) and Box-Cox Transformation (BCT) are widely used techniques that have demonstrated effectiveness across various fields. For instance, Bogner et al. employed the Normal Quantile Transform to normalize river flow data, thereby improving the accuracy and reliability of flood forecasting models [21]. In another study, Buchinsky applied Quantile Regression (QT) to analyze changes in returns to education and experience across different points of the wage distribution, as well as within-group wage inequality. The Box-Cox Transformation (BCT) was also incorporated in this analysis to appropriately transform the data for improved model fit [22]. Xie et al. used BCT as an early warning indicator for detecting abrupt climate changes [23], while Nagendra et al. utilized BCT to evaluate surface roughness in material science applications [24]. Similarly, Al Abbasi et al. applied QT to investigate the effects of rural transformation—specifically, high-value agricultural and non-agricultural employment—on income and poverty in Bangladesh [25].

Quantile transformation (QT) and Box-Cox transformation (BCT) improve prediction accuracy by regularizing the distribution of the dataset. In this study, QT and BCT methods are used to improve the accuracy of weight estimation. In literature, some studies that use these methods are summarized as follows: Peterson and Cavanaugh proposed a method that effectively transforms data into a normal distribution by introducing Ordered Quantile (ORQ) normalization. ORQ works consistently regardless of the underlying distribution and can be easily applied to new data. Its effectiveness is compared with other methods and the role of cross-validation in determining the best transformation is investigated. The technique was implemented on a car pricing dataset with the best Normalize R package [5]. Peng et al. highlighted the importance of transformations for mapping quantitative feature loci, and the empirical normal QT proved to be an effective method for normalizing feature values. In their study, they showed through extensive simulations that this transformation provides good control over power and type I error [6]. Rayner and MacGillivray studied the effects of Quantile-based methods in fitting g-and-k and adapted g-and-h distributions and showed that weighted methods perform better for small and medium-sized [7]. In another study, Hamzaoui et al. (2023) achieved a very high R^2 accuracy of 99.94% for different fish species by employing the SFI-XGBoost method, which combines VIF, Pearson correlation, and XGBoost [31]. Atkinson et al. (2021) examined the relationship between BCT power transformation and generalized linear models and proposed transformation models for positive and negative observations. In the study, normality, and variance homogeneity of the data were tried to be ensured by using methods that include the transformation of both sides. Zhang and Yang proposed new methods and algorithms for more efficient application of BCT on big data. This method aims to speed up transformation and model parameter estimation by scanning the data only once [8]. In Osborne's study, the relationship of the BCT transformation with traditional normalization transformations was discussed and how it was developed was discussed. Osborne stated that BCT offers better applications than traditional methods and provided examples of how this transformation can be applied using software such as SPSS and SAS [9].

In this study, fish weight estimation using ML methods using fish images is discussed. CatBoost, Random Forest, Polynomial Regression, and Support Vector Regression (SVR) models are used to accurately estimate fish weight. The success of each model was compared by applying different ML methods and to improve the accuracy, the improvements from the above QT and BCT studies were applied to the fish weight estimation project and a successful improvement was achieved. QT and BCT significantly improved the prediction accuracy by

significantly reducing the error rate in all models. The contributions of this study can be summarized as follows:

- To the best of our knowledge, Quantile Transformation (QT) and Box-Cox Transformation (BCT) have not previously been applied to the task of fish weight estimation. In this study, we incorporate these transformations into a novel fish weight estimation framework to evaluate their effectiveness.
- Machine learning methods are employed to estimate fish weight based on their morphological characteristics, and the error rates are further improved through the application of QT and BCT.
- Finally, we conducted a comparative analysis of weight estimation methods with and without transformations, evaluated the impact of the applied transformations, and assessed the improvements achieved.

The rest of the paper is organized as follows. The second section describes the dataset used and the evaluation metrics. In the third section, the methods applied in the preprocessing phase, machine learning models, and the experimental results and findings obtained from these models are presented in a comparative manner. Finally, the study ends with conclusions and future studies section.

Materials and Methods

In this study, CatBoost, Random Forest, Polynomial Regression and SVR machine learning methods were used to estimate weight from morphological features of fish and their performances were evaluated. In addition, the effects of Box-Cox Transformation (BCT) and Quantile Transformation (QT) on these methods were analyzed, and their impact on estimation performance was compared. The method with the lowest error rate was determined by making improvements and comparisons with different parameters. In total, 12 different models such as CatBoost, Random Forest, Polynomial Regression, SVR, and their versions with QT and BCT were tested. The results were analyzed among methods, especially for the effect of transformations. The dataset was divided into 80% training and 20% test data. The general progression of the steps followed in the study is presented in the method flowchart in Fig. 1. The figure the overall workflow of the study, which employs machine learning techniques to estimate fish weight based on morphological characteristics.

Dataset

This study utilizes the Fish Market dataset [10], which includes morphological measurements of fish. The dataset comprises 159 samples spanning seven species: Bream, Perch, Roach, Pike, Smelt, Parkki, and Whitefish. For each sample, seven morphological characteristics are recorded: Species, weight, length1, length2, length3, height and width. The weight estimation of fish is performed by utilizing various morphological characteristics. The meanings of these morphological features are given below and Fig. 2 shows the representation of these features on the fish. In this study, weight estimation was performed using the relationships and ratios between these features.

- Species: *Bream*, *Perch*, *Roach*, *Pike*, *Smelt*, *Parkki,* and *Whitefish*.
- Weight: Weight of the fish (in grams)
- Length1: Length from the nose to the beginning of the tail (in cm)
- Length2: Length from the nose to the notch of the tail (in cm)
- Length3: Length from the nose to the end of the tail (in cm)
- Height: % Maximal height as % of Length3Width: % Maximal width as % of Length3

Implementation Details

All experiments were conducted on a personal computer equipped with an Intel Core i5-13500HX CPU (2.5 GHz) and 16 GB of RAM.

Evaluation Metrics

In this study, Mean Absolute Error (MAE) and the Coefficient of Determination (R2) were used as evaluation metrics. MAE is defined as the average of the absolute differences between the predicted and actual values. Lower MAE values indicate that the model's predictions are closer to the true values. One of the key advantages of this metric is that it expresses the prediction error in the same unit as the target variable, which facilitates interpretability. Moreover, MAE is more robust to outliers compared to squared-error metrics, as it considers absolute deviations and does not disproportionately penalize large errors. In the literature, MAE is highlighted as a reliable and widely used metric for assessing regression performance [28]. The mathematical formulation of MAE is given in Equation (1).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

In this equation, n is the total number of data, yi indicates the actual values, and ŷi represents the predicted values. The R^2 metric was used to assess how well the independent variables in a regression model explain the variability of the dependent variable. It represents the proportion of variance in the dependent variable that can be predicted from the independent variables. An R^2 value of 1 indicates that the model explains the data perfectly, 0 indicates that it has no explanatory power, and negative values imply that the model performs worse than a simple mean-based prediction. In general, the closer the R^2 value is to 1, the stronger the explanatory power of the model. However, recent studies have emphasized that R^2, especially in linear regression, may sometimes overestimate the explained variance, which can lead to misleading interpretations [29]. The mathematical representation of R^2 is shown in Equation (2).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (2)$$

Within the scope of this study, the Box-Cox Transformation (BCT) was applied to adjust the distribution of variables toward normality. BCT is a parametric transformation method widely used for variance stabilization and approximation of normality in continuous and positive valued variables. The general formula of the transformation is shown in Equation (3).

$$y(\lambda) = \begin{cases} y^{\lambda} - 1, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases} \quad (3)$$

here, y is the original data values to be transformed, λ is a parameter that determines the intensity and direction of the transformation. The λ values used in the study were determined by optimizing with the maximum log-likelihood method. In this method, the log-likelihood function for the transformed data set is as follows.

$$\log L(\lambda) = \frac{\pi}{2}\log(\mathrm{SSE})(\lambda) \quad (4)$$

where, SSE(λ) represents the sum of squared errors of the transformed data and n represents the number of observations. The parameter λ, which provides the maximum log-likelihood value, is used as the optimal transformation coefficient that ensures that the data set is distributed closest to normality. This method determines the transformation parameters systematically and objectively, increasing the reproducibility and statistical reliability of the analysis results.

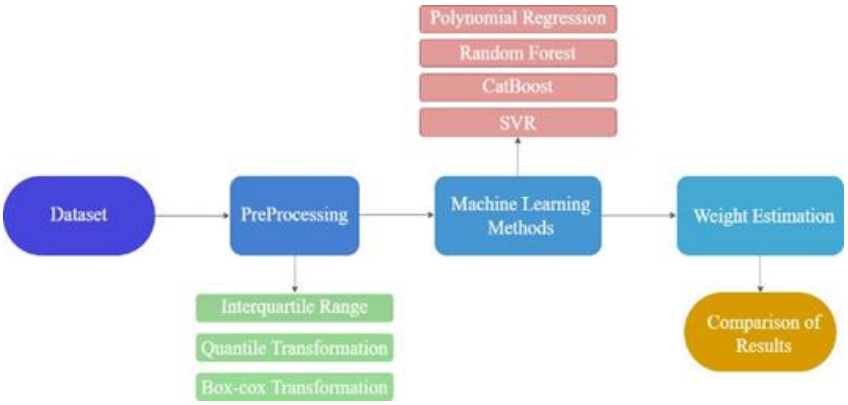


Figure 1. Flowchart of the proposed method.

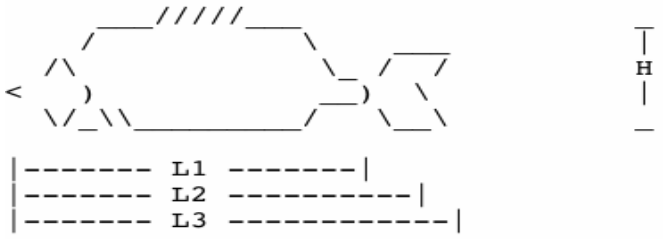


Figure 2. Illustration of the morphological features measured on fish, including Length1 (L1), Length2 (L2), Length3 (L3), Height (H), and Width (W) [11].

Experimental Results and Findings

This section outlines three key aspects: changes in preprocessing steps before and after applying QT and BCT, performance variations of machine learning algorithms under each condition, and the effect of these transformations on prediction outputs.

Preprocessing Steps

Several data-related challenges hinder the reliable and effective performance of machine learning and data analysis processes. These problems can be identified and corrected with the data preprocessing step. In the data preprocessing phase, the relationships of morphological features with each other, especially with weight, are analyzed, outliers are identified, and missing data are removed. When the data preprocessing phase is skipped, the model's learning process is adversely affected, prediction accuracy decreases, the model tends to learn biasedly, and its generalization ability may be weakened, leading to erroneous results. Data preprocessing steps should be implemented to reduce the impact of such problems.

In this study, all preprocessing steps are kept identical except for the QT, BCT, and method components. This approach allows for a more accurate comparison of

prediction errors and isolates the effects of the QT and BCT transformations on weight estimation.

Attribute Representation

To analyze the value distribution of each morphological feature in the dataset, bar charts were generated. Fig. 3 presents the column plots for species, weight, width, height, length1, length2, and length3. In the distribution of these graphs, species are clustered in certain index ranges, indicating categorical distinctions in the dataset. Fig. 3(a) shows the species distribution in the dataset. In Fig. 3(b), the weight variable shows sharp increases and decreases in certain index ranges, indicating that some fish species are much heavier. The width and height plots in Fig. 3(c) and Fig. 3(d) show that different species are more dominant in certain ranges, while the length variables (length1, length2, length3) in Fig. 3(e), Fig. 3(f) and Fig. 3(g) show similar distributional trends and increase on the horizontal axis. Overall, the variables exhibit different distributions within specific index ranges, indicating that the dataset has distinct clustering by species, reflecting measurable differences between fish species. The fact that fish exhibit different distributions within specific index ranges directly affects weight estimation, requiring consideration of species-based differences. Since each fish species has its weight range, their relationships with measurements such as length, width, and height also differ. When the graphs are analyzed, it is seen that some variables show a right-skewed (positive skewness) distribution. The distribution of weight and height variables exhibits a right-skewed character with a long tail structure extending from small to large values. Length1, length2, and length3 variables are skewed to the right with a distribution of low values and decreasing towards large values. The effect was observed when QT and BCT were applied to remove these skews.

Quantile Transformation (QT)

QT is a pre-processing step used to transform distributions of data. Instead of creating a new dataset, this transformation creates a transformed version of the original dataset. In this transformation, the data is rescaled according to a given probability distribution. It is especially preferred for data with different distributions to achieve a more homogeneous distribution. QT is usually performed by placing the order of each observation into the corresponding Quantile in the target distribution.

There are two common approaches to this transformation: uniform distribution and normal distribution. The uniform distribution is a type of distribution where each value is chosen with equal probability, and with this method, the data is redistributed with equal probabilities between 0 and 1. In this distribution, the data only lie between 0 and 1 and do not take negative values. In the normal distribution, the data are symmetrically distributed around the mean and can also take negative values. The uniform distribution was used in the study. This process ensures that the values are evenly distributed between 0 and 1, without any negative values. Furthermore, by leveraging the Box-Cox Transformation's ability to approximate a normal distribution, the study aims to compare the results obtained through this alternative transformation approach. QT has been applied to all columns except the Species column. The other columns were rearranged according to a specific probability distribution and subjected to QT to ensure a homogeneous distribution of the data. Fig. 4 shows the original distribution of the dataset and its distribution via QT. As shown in the figure, the dataset exhibits a more homogeneous distribution after the application of QT.

Before the QT transformation, the data exhibited substantial right skewness, with most values clustered near the lower end and a long tail extending toward higher values. To address this issue, QT was applied, transforming the data into a uniform distribution. After the transformation, the values were evenly distributed within the [0, 1] range, reducing skewness and the influence of outliers, and producing a more balanced dataset suitable for machine learning models.

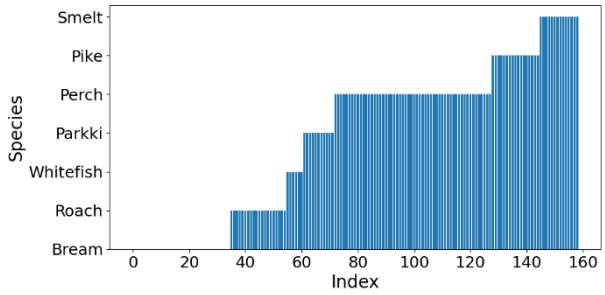
Box-Cox Transformation (BCT)

In regression analysis, the accuracy of most models increases when the data resemble a normal distribution. BCT is a mathematical method used to bring the distribution of the data closer to a normal distribution. This transformation attempts to transform the data into a normal distribution by applying functions such as a logarithm or square root to the data. BCT is widely used to improve the accuracy of the model, especially when working with volatile and skewed data, the Fish Market dataset is skewed to the right as can be seen in the bar charts in Fig. 3. This transformation aims to flatten the distribution of the data, make it symmetric, and perform a better regression analysis. BCT does not work with negative values, since there are no negative values in the Fish Market dataset, it is suitable for BCT. Fig. 5 shows the original dataset distribution, the distribution of the dataset after BCT, and the Kernel Density Estimation (KDE) curve of this distribution. By obtaining a continuous probability density function with the KDE curve, it can be observed in which intervals the data is concentrated. The graphs indicate that the data distribution after BCT is closer to the normal distribution.

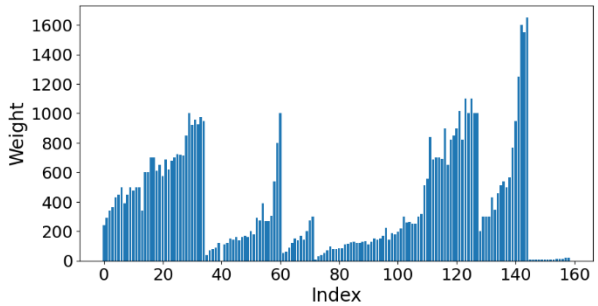
BCT proved highly effective in converting the right-skewed and potentially multimodal distributions into more symmetric, near-normal forms. This transformation serves as a critical preprocessing step for enhancing the performance of parametric statistical tests and machine learning models. Additionally, BCT alters the scale of the data, making it more compact and facilitating more stable and interpretable model training.

Interquartile Range (IQR)

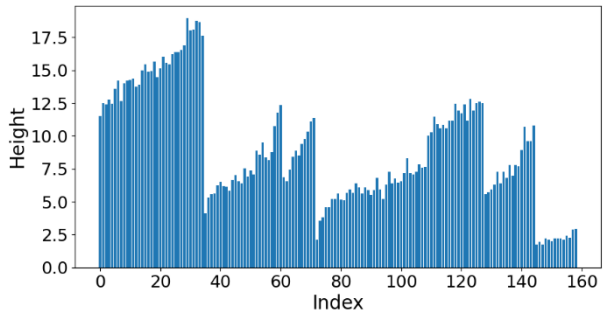
The IQR method is a statistical method used to identify outliers in a data set. This method detects and visualizes unusually low or high values based on the difference between the 25% quartile of Q1 and the 75% quartile of Q3. To identify outliers, a threshold of 1.5 times the difference between Q1 and Q3 was set. In this way, outliers in the data



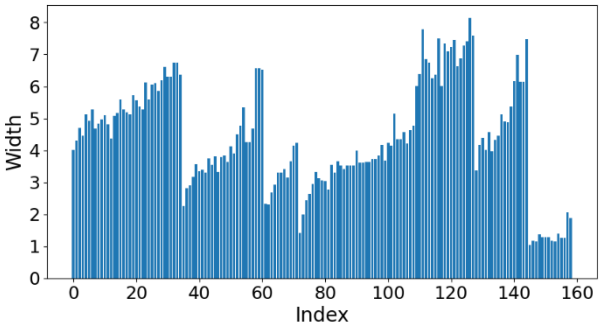
(a) Species



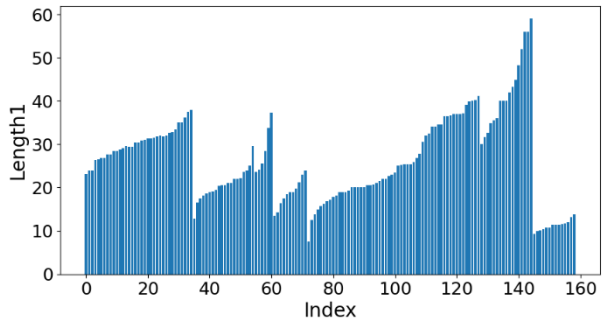
(b) Weight



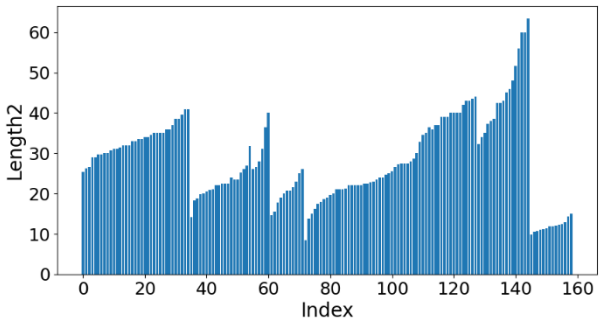
(c) Height



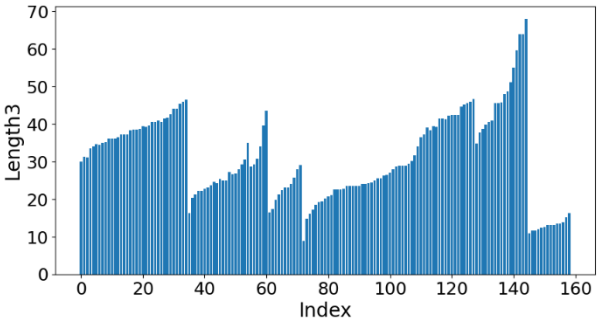
(d) Width



(e) Length1



(f) Length2



(g) Length3

Figure 3. Distribution of morphological features in the dataset.

set that could affect the analysis results were identified, resulting in more reliable and consistent results. The lower bound is shown in Equation (4) and the upper bound is shown in Equation (5).

$$L_b = \text{Q1} - 1.5 * \text{IQR} \quad (4)$$

$$U_b = \text{Q3} + 1.5 * \text{IQR} \quad (5)$$

Fig. 6(a) shows the plot of outliers detected by the IQR method before the transformation is applied, there are outliers because the raw data has a skewed distribution. Fig. 6(b) presents the results of the outliers detected with the IQR method after QT, while Fig. 6(c) shows the effect of the IQR method on the outliers after BCT. When the graphs were analyzed, outliers were detected in length1, length2, and length3 before the transformation. However, QT and BCT eliminate these outliers.

Correlation Matrix

The correlation matrix is used to analyze and interpret the relationships between variables. It is a matrix showing the linear relationship between variables. It is used in statistics and data analysis processes to understand how variables are related to each other. This matrix contains the correlation coefficients between pairs of variables. This correlation coefficient takes values between -1 and 1. Negative values mean that there is a negative correlation between the variables, i.e. one increases while the other decreases, while positive values mean that there is a positive correlation between the variables, i.e. one increases while the other increases or one decreases while the other decreases. 0 indicates that there is no linear relationship between them. In the scope of this study, we observed the correlation between morphological features in the absence of transformations, the correlation between morphological features after QT, and the correlation between morphological features after BCT, and how transformations affect the relationships between features. Fig. 7(a) shows the correlation matrix without transformation, Fig. 7(b) shows the correlation matrix after QT and Fig. 7(c) shows the changes in the correlation matrix after BCT. When the correlation matrices are analyzed, generally positive correlations are detected between the features before any transformation is applied. In particular, a strong positive correlation is observed among the features Length1, Length2, and Length3. This suggests that these features are highly interrelated and likely convey overlapping or redundant information. After applying QT and BCT, the correlation between the features generally increased at similar rates. However, this increase was found to be more pronounced in QT. This suggests that QT reveals the relationships between features more strongly and may be more effective in the process of normalizing the data distribution.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used to identify the components that explain the highest variability in the data. Given the relationships in the correlation matrix, applying dimensionality reduction techniques on highly correlated features can minimize the risk of overlearning by reducing the complexity of the model. Moreover, this approach can contribute to improved model generalizability and reduced error rates, thereby enhancing overall model performance. Fig. 8(a) shows the variance explained for each principal component in the absence of transformation, Fig. 8(b) shows the variance explained for each principal component after QT and Fig. 8(c) shows the changes in the variance explained for each principal component after BCT.

These graphs show that in the absence of transformation, the first principal component explains a large proportion of the total variance. This shows that the data is highly correlated and PCA can represent a large portion of the data with a single component. After the transformations, the transformations provide a more even distribution of variance, increasing the contribution of the second and third components, but the first component is still dominant.

Machine Learning Methods

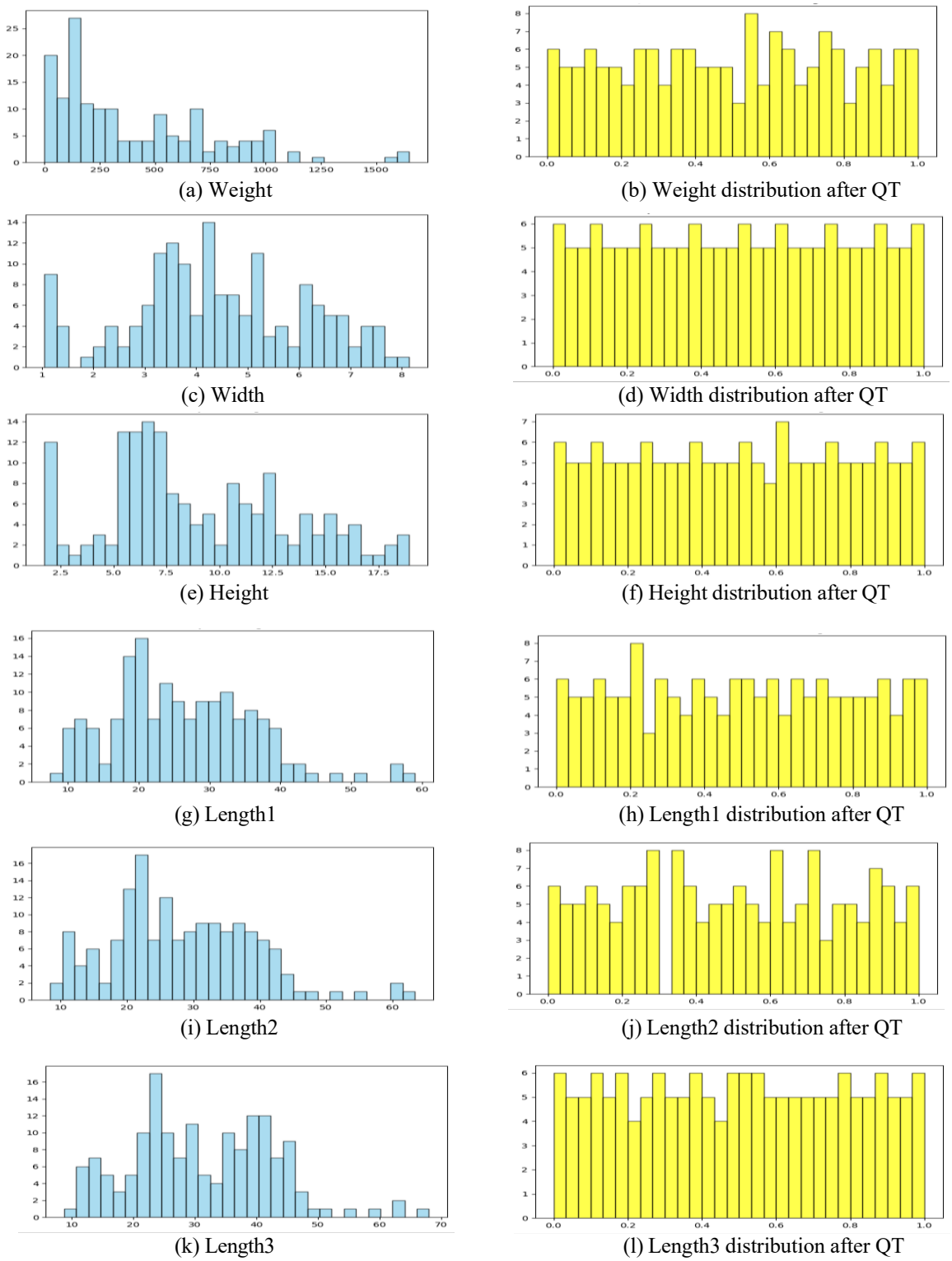
In studies involving structured datasets with multiple features, machine learning (ML) approaches offer powerful alternatives to traditional statistical methods, particularly when developing prediction models using high-dimensional data [12]. In this study, the performance of four ML methods including, CatBoost, Random Forest, Polynomial Regression, and Support Vector Regression (SVR) is compared across three data versions: untransformed, QT-transformed, and BCT-transformed. The results and the impact of these transformations on model performance are discussed in detail.

CatBoost

CatBoost has emerged as a powerful tool for ML tasks involving big data [13]. It is a machine learning algorithm belonging to the Gradient Incremental Decision Trees (GADT) family. In this study, it was used for a small dataset. CatBoost can be used in both classification and regression problems. CatBoost Regression is the version of CatBoost used in weight estimation problems. One of the most important features of CatBoost is that it can work directly with categorical variables and process these variables effectively. In the Fish Market dataset, the species column is a categorical variable. CatBoost applies several methods to deal with categorical features. For one-hot encoded features, no special processing is required, and the histogram-based approach used for partition search can be easily adapted to this case [14]. When using CatBoost, there is no need to convert this column into numerical data. CatBoost is a method that aims to build a strong model by successively building weak models, usually decision trees. At each step, it focuses on correcting the prediction errors of the previous model. The process starts with a simple model, and each new tree is trained to minimize the errors of the previous model over gradients. This process improves the performance of the model step by step. The training process of the CatBoost model is shown in Fig. 9(a) before applying the transformation, Fig. 9(b) after QT, and Figure 9 (c) after applying BCT. In the graphs, it is observed that

the error value in the test data decreases more than the training data. MAE values decrease significantly after QT and BCT. The lowest MAE value is obtained after Quantile transformation as seen in the graphs. The error values of the training data show a better reduction than the error values of the test data. However, as the model becomes more complex, the risk of overfitting arises.

A lower error was observed in the train data compared to the validation data, and this was evaluated in terms of the possibility of overfitting. However, after hyper parameter optimization and multi-level validation, the difference between training and validation errors decreased, which



(a) Weight (b) Weight distribution after QT

(c) Width (d) Width distribution after QT

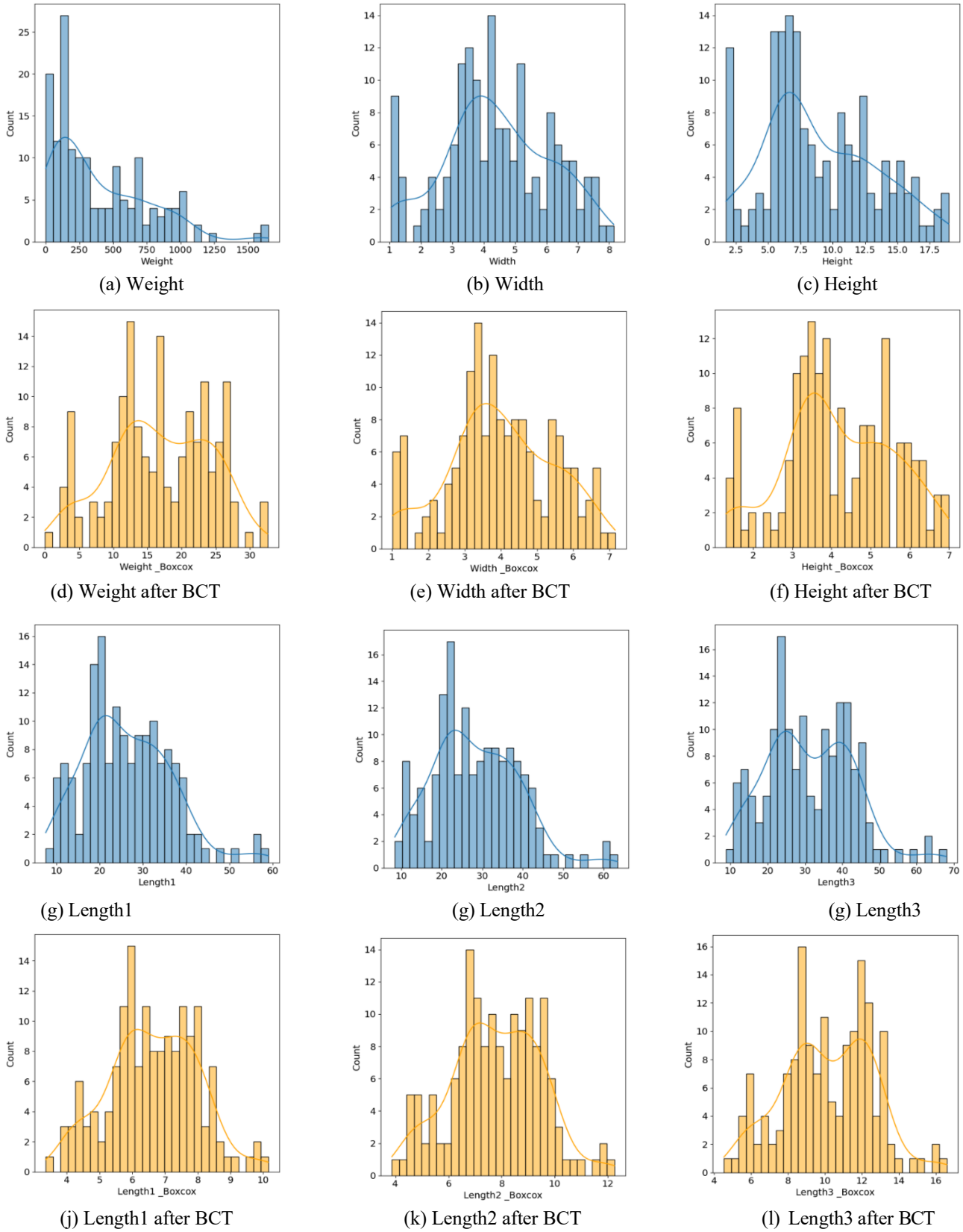
(e) Height (f) Height distribution after QT

(g) Length1 (h) Length1 distribution after QT

(i) Length2 (j) Length2 distribution after QT

(k) Length3 (l) Length3 distribution after QT

Figure 4. Original and Quantile Transformed (QT) distributions of fish morphological features: (a)-(l) show before and after QT transformations for Weight, Width, Height, Length1, Length2, and Length3, respectively.



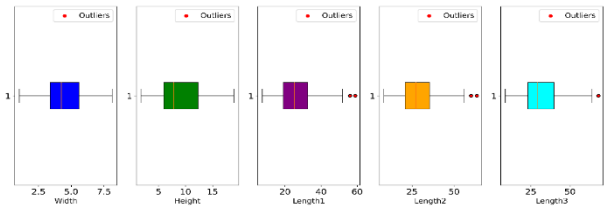
(a) Weight (b) Width (c) Height

(d) Weight after BCT (e) Width after BCT (f) Height after BCT

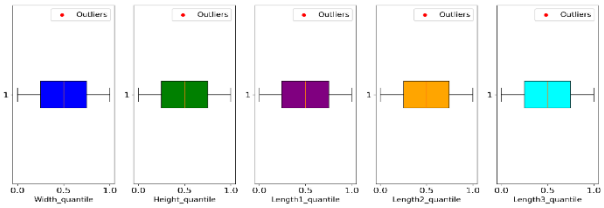
(g) Length1 (g) Length2 (g) Length3

(j) Length1 after BCT (k) Length2 after BCT (l) Length3 after BCT

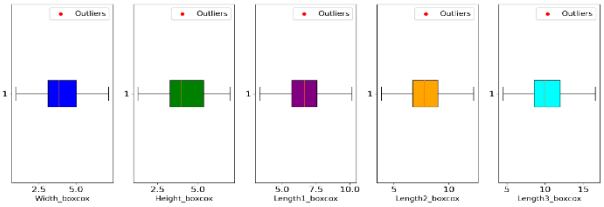
Figure 5. Original and Box-Cox Transformed (BCT) distributions of fish morphological features with Kernel Density Estimation (KDE) curves: (a)-(l) show before and after BCT transformations for Weight, Width, Height, Length1, Length2, and Length3, respectively



(a) Outliers detected by IQR method in the absence of transformation

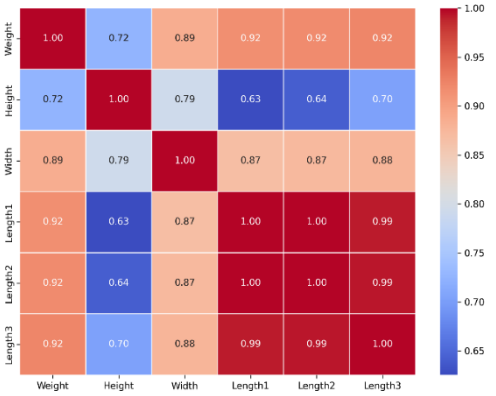


(b) Outliers detected by IQR method after QT

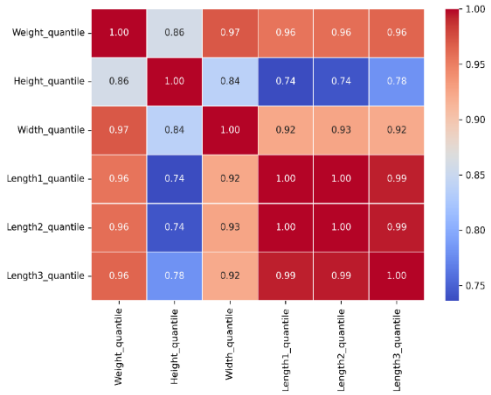


(c) Outliers detected by IQR after BCT

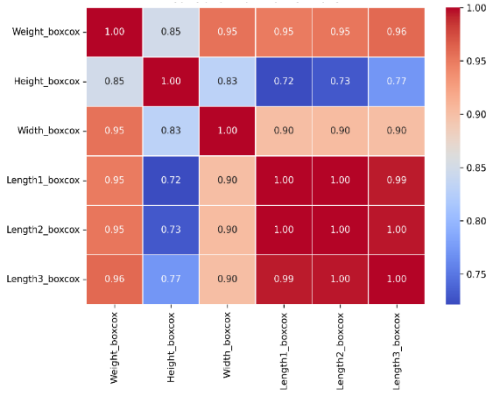
Figure 6. Outliers detected by IQR in the absence of conversion, after QT and after BCT.



(a) Correlation matrix without transformation



(b) Correlation matrix after QT



(c) Correlation matrix after BCT

Figure 7. Correlation matrices of morphological features.

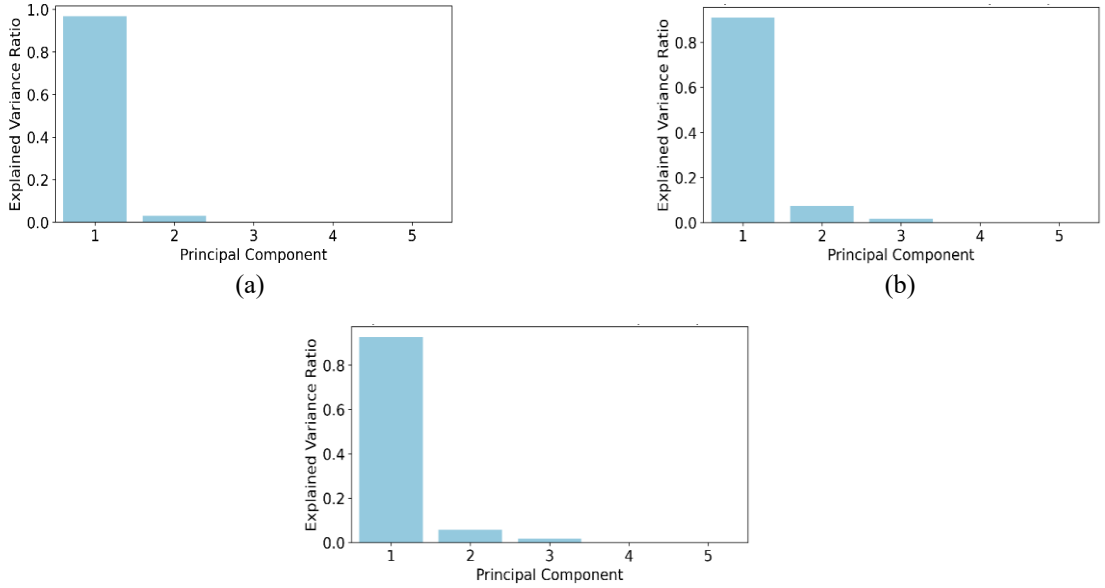


Figure 8. Variance explained by principal components: (a) Without transformation, (b) After QT, (c) After BCT.

reduced the risk of overfitting and improved the model's generalization performance.

CatBoost

Grid search is used to improve the performance of the model for hyper parameter optimization and to determine the best combination of hyper parameters. Using cross-validation with Grid search, different combinations of iterations, learning rate, depth and fold count hyper parameters were tested, and the MAE values of the combinations without transformation in Fig. 10(a), after QT in Fig. 10(b) and after BCT in Fig. 10(c) are shown graphically. In this process, the performance of the model was evaluated according to the MAE metric, and the hyper-

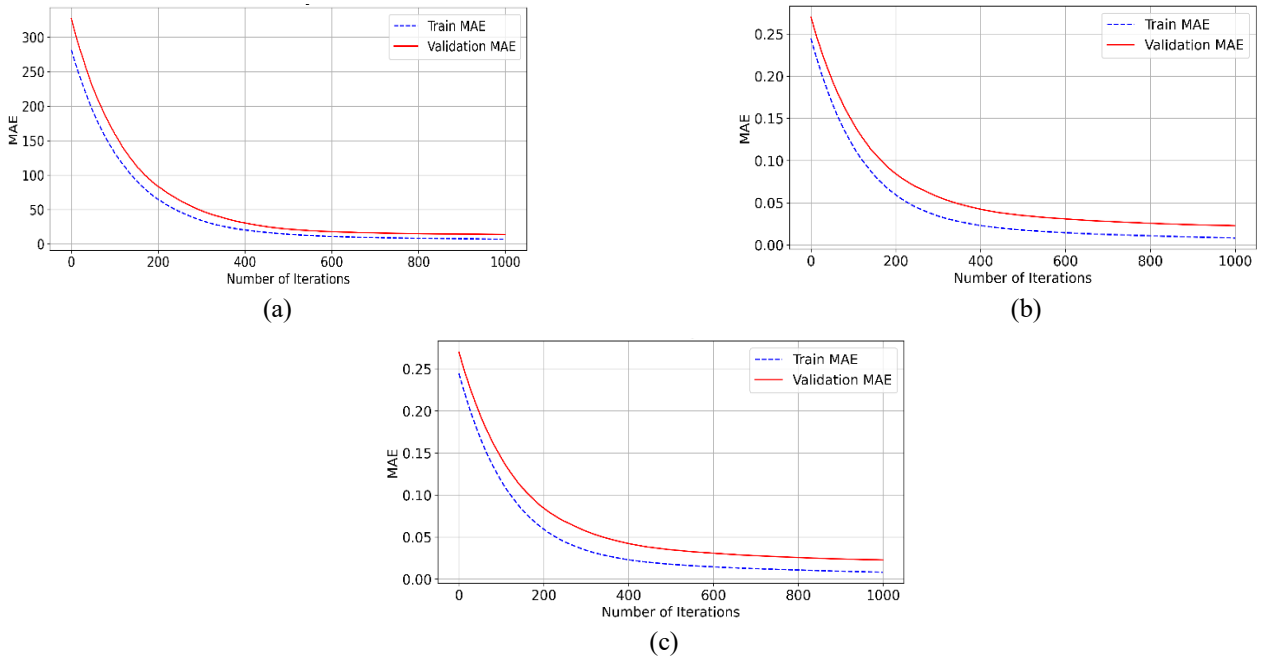


Figure 9. CatBoost model training process graphs: (a) Without transformation, (b) After QT, (c) After BCT.

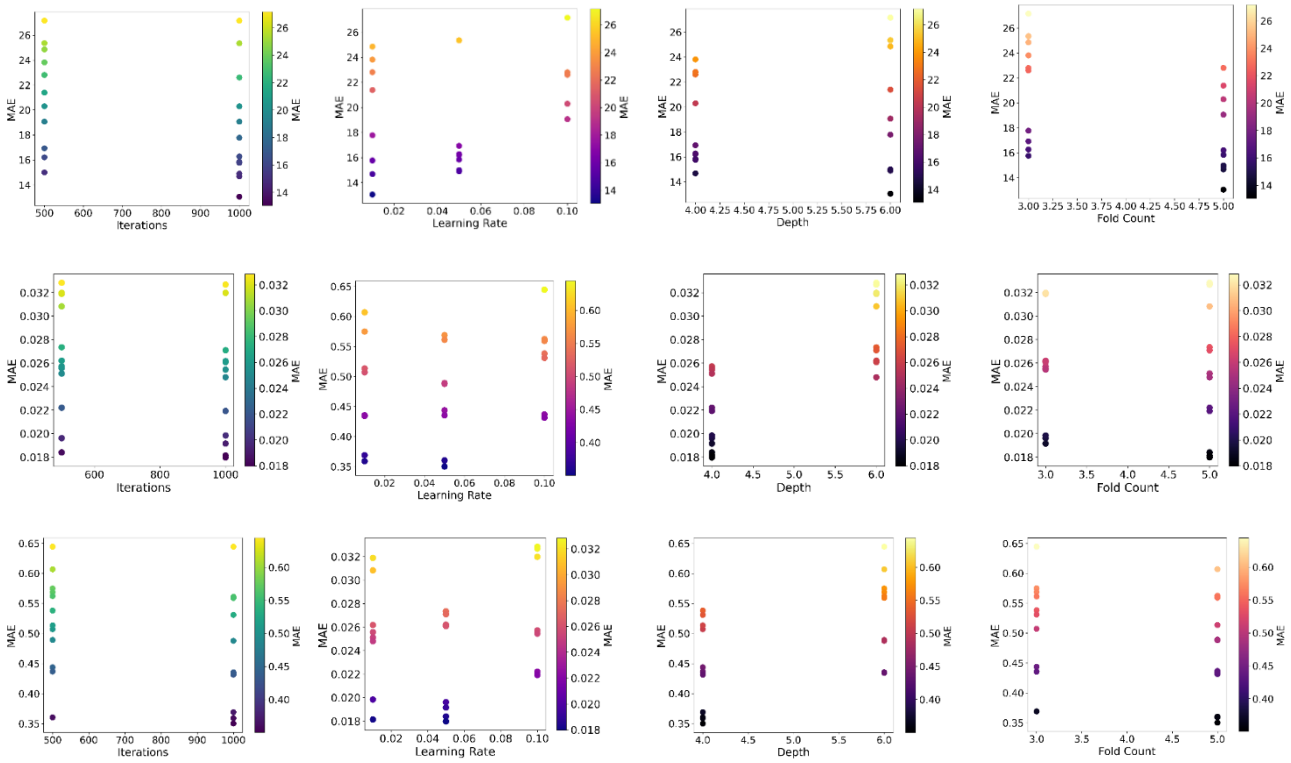


Figure 10. Grid search results showing MAE values for different combinations of CatBoost hyperparameters (iterations, learning rate, depth, and number of layers). The first column corresponds to the original data (untransformed), the second column to the data with QT applied, and the third column to the data with BCT applied.

parameters with the lowest MAE values without transformation, after QT and after BCT were selected. Thus, it was aimed to generalize the model better and to prevent overlearning and overfitting. Values giving the lowest error rate in the absence of transformation: iterations: 1000, learning rate: 0.01, depth: 6, fold count: 5. Values giving the least error rate after QT: iterations: 1000, learning rate: 0.05, depth: 4, fold count: 5 Values giving the least error rate after BCT: iterations: 1000, learning rate: 0.01, depth: 4, fold count:3. These values are different due to the transformation of the dataset after the transformations.

Random Forest

Random Forest is an ML method that can be used in both classification and weight estimation models. One study examines the historical development of Random Forest, its successful applications, and comparisons with other classifiers [15]. This method is used for a regression problem and compared with other ML methods in this study. It is an ensemble of decision trees, generates multiple decision trees, and makes a final prediction by taking predictions from these trees and averaging these predictions. This method uses the bagging technique to reduce overfitting and increase the model's generalization ability. This technique allows the model to produce more balanced and reliable results by training each decision tree on different subsets of data. It cannot work with categorical data, so categorical data was converted into numerical data using the label encoder method.

Polynomial Regression

Polynomial regression is a modeling method used to understand the relationship between independent and dependent variables. It is a preferred ML method when the link between variables is not linear.

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n + \epsilon \qquad (5)$$

here, b_0, b_1, b_2, b_n are the coefficients of the independent variable x, x^2 is its square, ϵ is the error term. In polynomial regression, more complex and flexible models were created by adding higher-order terms of the independent variable. For example, fish length (x) as an independent variable presented a certain linear relationship when included in the model alone, while the square of length (x^2) helped to better capture growth trends and make weight estimation more accurate. Although this method allows for more successful modeling of nonlinear relationships, the optimum model complexity was determined to avoid overfitting, as polynomials of very high degree can lead to overfitting.

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning method commonly applied to regression problems and has been utilized across various domains. For instance, a study

by [27] examined treatment processes and the management of patients who achieved Sustained Virologic Response (SVR), while [26] explored the application of hybrid machine learning techniques, particularly SVR, for wind forecasting and the management of ramp events. In this study, SVR is employed for the estimation of fish weight. SVR is an effective method for regression problems with non-linear relationships such as fish weight estimation. Its goal is to find the best hyperplane by ignoring the data within the specified error tolerance (ε). It builds models with support vectors only, thus avoiding overlearning and improving generalization performance. Non-linear relationships can be captured by using morphological characteristics of the fish as independent variables and weight as dependent variable. In this model, the Grid Search method was used in hyper parameter optimization to improve the performance of the model and to determine the most appropriate hyper parameter combination. C, epsilon, and gamma values were determined by the Grid search method for non- transformed, after QT and after BCT. These parameters significantly affect the learning process and performance of the SVR model. The C parameter determines the error tolerance of the model. A high C value leads to a tighter fit of the model to the training data, but increases the risk of overfitting. A low C value creates a more flexible model, but may not fit the training data perfectly. The epsilon parameter determines the acceptable amount of error in the model's predictions. A small epsilon value produces more accurate predictions, while a large epsilon value produces a more general model and accepts more errors. The gamma parameter determines the complexity of the kernel function. When 'scale' is selected, the gamma value is calculated automatically according to the characteristics of the data. When 'auto' is selected, the gamma is set to a fixed value. Setting these hyper parameters correctly improves the overall performance of the model, leading to more reliable and accurate predictions and reducing the risk of overfitting. Fig. 11(a) shows the MAE vs. MAE values of the combinations of “C”, “Epsilon” and “Gamma” parameters in the absence of transformation, in Fig. 11(b) after QT and in Fig. 11(c) after BCT. 'C' in Fig. 11(a): 1000, 'epsilon': 1, 'gamma': 'scale' and 'C': 100, 'epsilon': 0.01, 'gamma': 'auto' in Fig. 11(b): 100, 'epsilon': 0.01, 'gamma': 'auto' in Fig. 11(b), and 'C': 1000, 'epsilon': 0.01, 'gamma': 'scale' gives the lowest MAE value.

Quantitative Comparison

The values obtained with Catboost, Random Forest, Polynomial Regression and SVR methods without transformation, after QT, and after BCT are given in Table 1. As shown in Table, among the models applied without transformation, CatBoost yielded the lowest Mean Absolute Error (MAE), with a value of 14.0020. The highest MAE value was obtained in the SVR model with a value of 45.49. In general, it is observed that the average absolute error in the models decreases significantly after QT and BCT. This shows that the transformations significantly reduce the error rate and increase the success of the model. There is a greater reduction in the error value after QT than BCT, indicating that QT gives better results for the dataset. The CatBoost+QT experiment had the lowest average absolute error in these 12 runs. In the R^2 values, the highest value is obtained in the SVR + BCT method, which means that the model explains all the data almost perfectly, and the R^2 values are generally close to 1, which shows that it successfully explains a large part of the data in all methods and provides a good fit. Fig. 12 shows the boxplot of weight prediction with CatBoost model without transformation, weight prediction with CatBoost model after QT and weight prediction with CatBoost model after BCT. Fig. 13 shows the box plot of weight estimation with the Random Forest model after QT, and weight estimation with the Random Forest model after BCT. Fig. 14 shows the box plot of weight prediction with the Polynomial Regression model without transformation, weight prediction with the Polynomial Regression model after QT, and weight prediction with the Polynomial Regression model after BCT. Fig. 15 shows the box plot of weight prediction with the SVR model without transformation, after QT, and BCT, respectively.

An analysis of Figs. 12 through 15 clearly demonstrates the impact of the applied transformations on the distribution of the data. In the graphs without transformation, it is seen that the weight values are distributed in a wide range and the data set exhibits a high degree of variability. This shows that the data distribution has a skewed structure, especially with the effect of extreme values. In the graphs obtained after QT, the distribution is made more symmetric by scaling the weight values between 0 and 1. This transformation significantly reduced the effect of outliers and increased the homogeneity of the data. In graphs obtained after BCT, the weight values were scaled between 0 and 30, closer to their original units. This transformation reduced, but did not completely eliminate, the effect of outliers by bringing the data closer to a normal distribution.

Limitations of the proposed work

The Fish Market dataset used in this study is one of the rare open-access resources that provides both morphological features of different fish species and their corresponding weight measurements. In the literature, there is a limited number of reliable datasets that include both physical measurements and validated weight information, making this dataset the most suitable option for our study. However, the dataset consists of only 159 samples, which limits the generalization capability of the applied machine-learning algorithms. Additionally, due to the imbalanced representation of certain species, the models may perform poorly on underrepresented classes. The small sample size also increases the risk of overfitting, where the model performs well on training data but fails to produce accurate predictions on unseen data.

To address these limitations, future work will focus on constructing larger datasets with more representative samples, thereby improving both the accuracy and generalizability of the models. In addition, synthetic data generation techniques based on the statistical characteristics

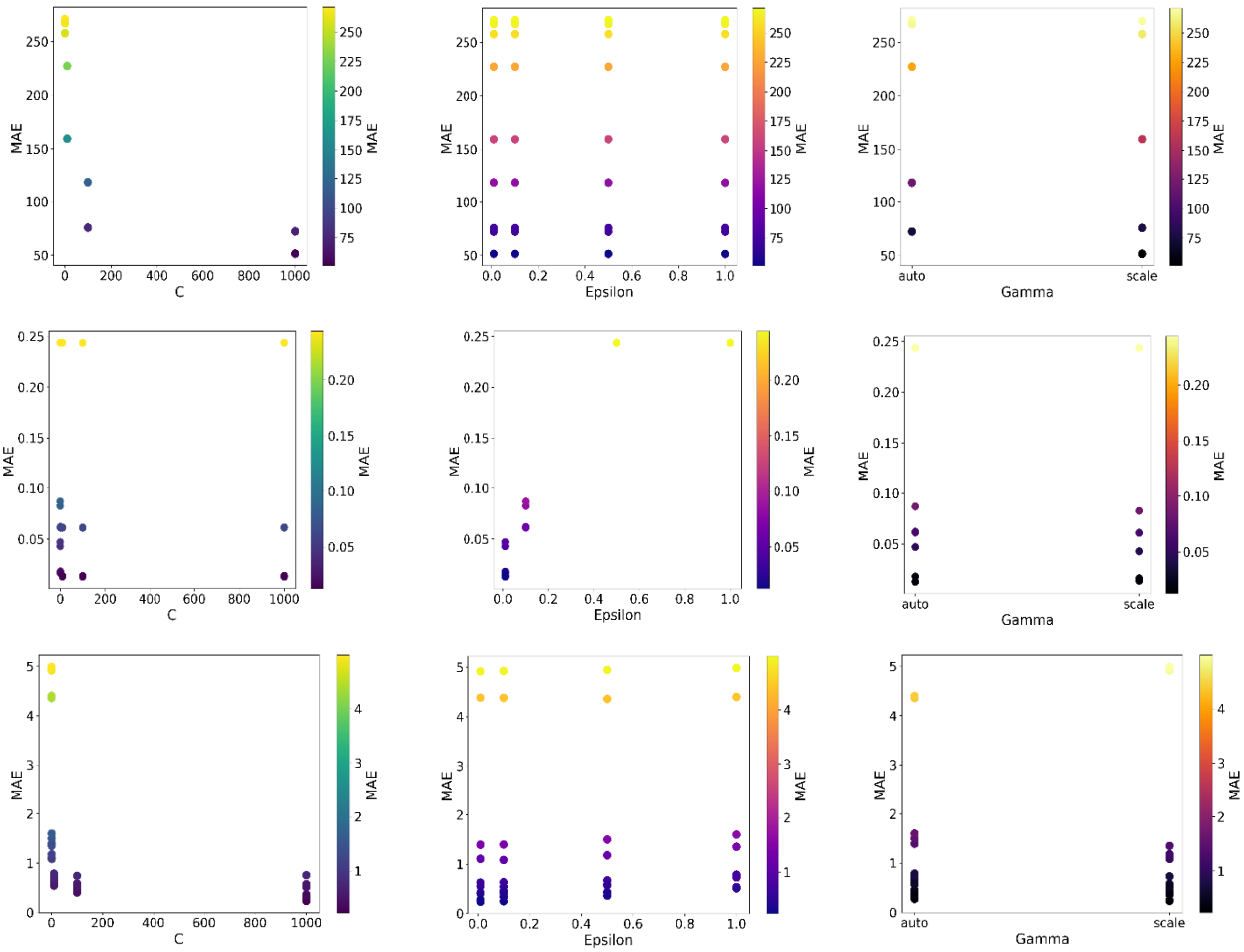


Figure 11. MAE values for Support Vector Regression (SVR) with different combinations of hyper parameters (C, epsilon, gamma). The first row corresponds to the original data (no transformation), the second row to QT-applied data, and the third row to BCT-applied data.

Table 1. Performance Comparison of Models with and Without Data Transformations

Method	MAE	R^2
CatBoost	14.0020	0.9958
CatBoost + QT	0.0171	0.9947
CatBoost + BCT	0.3302	0.9971
Random Forest	44.53	0.9692
Random Forest + QT	0.03	0.9815
Random Forest + BCT	0.77	0.9834
Polynomial Regression	42.8059	0.9732
Polynomial Regression + QT	0.0306	0.9847
Polynomial Regression+ BCT	0.6533	0.9882
SVR	45.49	0.9622
SVR + QT	0.01	0.9981
SVR + BCT	0.22	0.9986

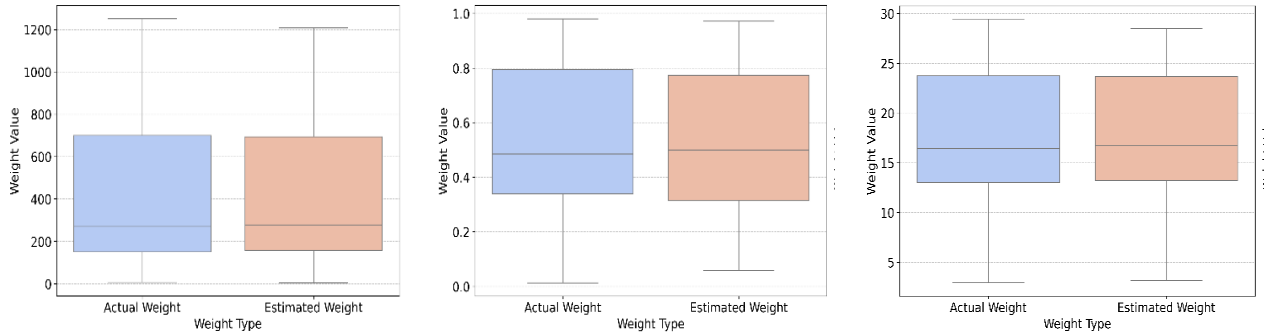


Figure 12. Box plot comparison of actual and predicted fish weight using CatBoost. From left to right, the columns correspond to the original data (no transformation), QT-applied data, and BCT-applied data.

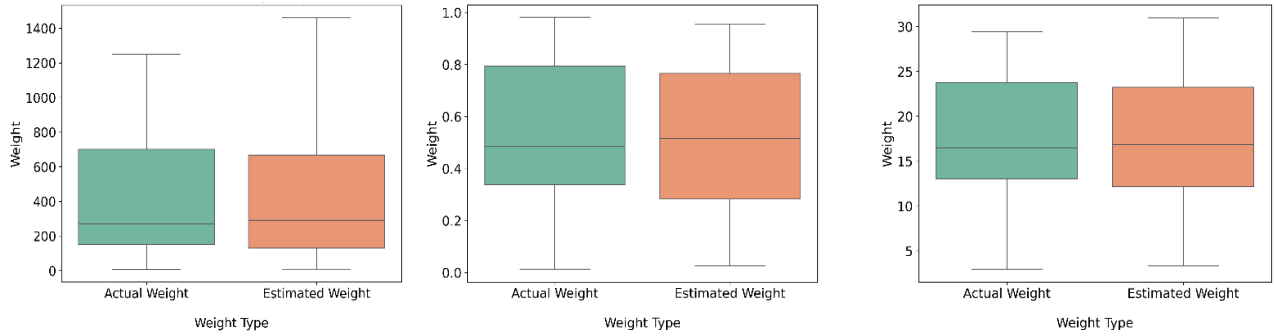


Figure 13. Box plot comparison of actual and predicted fish weight using Random Forest. From left to right, the columns correspond to the original data (no transformation), QT-applied data, and BCT-applied data.

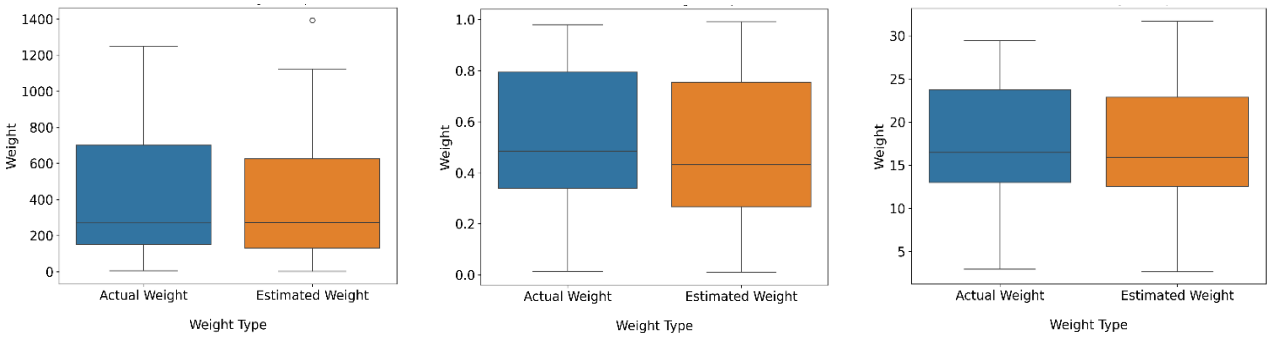


Figure 14. Box plot comparison of actual and predicted fish weight using Polynomial Regression. From left to right, the columns correspond to the original data (no transformation), QT-applied data, and BCT-applied data.

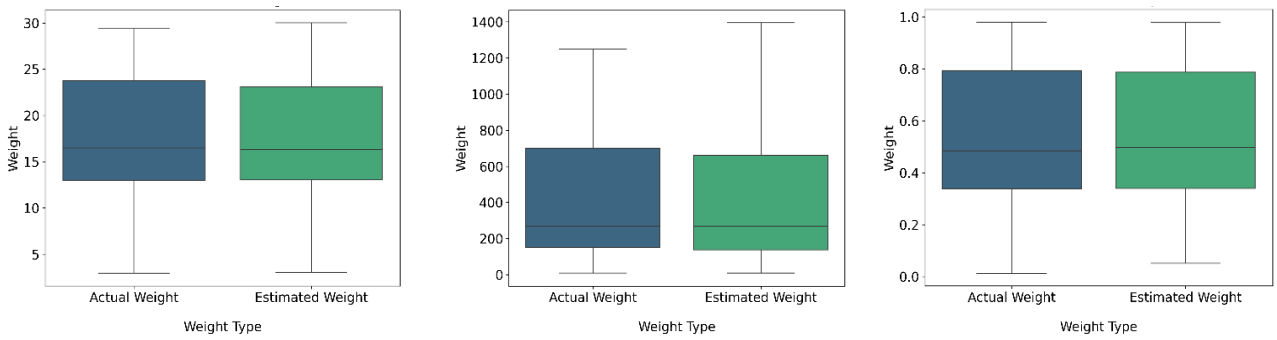


Figure 15. Box plot comparison of actual and predicted fish weight using SVR. From left to right, the columns correspond to the original data (no transformation), QT-applied data, and BCT-applied data.

of the existing data may be explored to enhance data diversity. If visual data is incorporated in subsequent phases, image-based data augmentation techniques could also be employed to improve model robustness. Expanding the dataset in these ways will significantly enhance the statistical reliability of the findings and their potential applicability in real-world scenarios.

Conclusions and Future Works

This study investigates the effects of data transformation methods on improving the fish weight estimation model's accuracy. The QT and BCT transformation methods applied to the input data as a preprocessing step appear to significantly reduce weight estimation errors by reconstructing the fish weight distribution closer to normal form. These transformations, which are applied as a preprocessing step, correct the skewed distribution in the dataset and enable machine learning models such as CatBoost, Random Forest, Polynomial Regression, and SVR to generalize better and perform better than the weight estimation results obtained without preprocessing on the raw dataset. Among the QT and BCT transformation methods, QT outperforms BCT by giving the lowest MAE value in all machine learning models. QT applied on the original dataset dramatically improves by reducing the MAE value from 14.002 to 0.017 with the CatBoost model, which gives the best accuracy, while BCT achieves an MAE value of 0.330 in the same model. The experimental results demonstrate that not only machine learning model selection has a significant impact on accuracy, but also a well-chosen transformation has a significant impact on error reduction.

The findings show that proper transformation and data normalization are important factors in maximizing model prediction performance. Future studies aim to learn more complex features from image or video frames using advanced deep learning architectures such as Convolutional Neural Networks (CNNs) for fish weight estimation and to make real-time fish weight estimation using camera and sensor-based technologies. Moreover, the model's performance on fish weight estimation using many parameters will be examined with multi-model data fusion, such as the use of image data together with sensor readings on water quality, temperature, and sonar- based measurements of fish size.

Ethics committee approval and conflict of interest statement

There is no conflict of interest with any person / institution in the article prepared.

Authors' Contributions

Esen H: Methodology, Software, Visualization, Data extraction, Writing – original draft.

Taşdelen H: Methodology, Data extraction, Writing – original draft.

Küçük S: Conceptualization, Data curation, Methodology, Validation, Supervision, Writing – review & editing.

Aksakallı IK: Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing.

All authors have read and agreed to the published version of the manuscript.

References

[1] S. B. Gutzmann, E. E. Hodgson, D. Braun, J. W. Moore, and R. A. Hovel, "Predicting fish weight using photographic image analysis: a case study of broad whitefish in the lower Mackenzie River watershed," *Arct. Sci.*, vol. 8, no. 4, pp. 1356–1361, Dec. 2022. Accessed on: Jun. 30, 2025. doi: 10.1139/AS-2021-0017.

[2] D. A. Konovalov, A. Saleh, D. B. Efremova, J. A. Domingos, and D. R. Jerry, "Automatic weight estimation of harvested fish from images," in *Proc. 2019 Digital Image Computing: Techniques and Applications (DICTA)*, Dec. 2019, doi: 10.1109/DICTA47822.2019.8945971.

[3] R. Islamadina, N. Pramita, F. Arnia, and K. Munadi, "Estimating fish weight based on visual captured," in *Proc. 2018 Int. Conf. Inf. Commun. Technol. (ICOIACT)*, vol. 2018-January, pp. 366–372, Apr. 2018, doi: 10.1109/ICOIACT.2018.8350762.

[4] S. Suwannakhun and P. Daungmala, "Estimating pig weight with digital image processing using deep learning," in *Proc. 14th Int. Conf. Signal Image Technol. Internet Based Syst. (SITIS)*, pp. 320–326, Jul. 2018, doi: 10.1109/SITIS.2018.00056.

[5] R. A. Peterson and J. E. Cavanaugh, "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era," *J. Appl. Stat.*, vol. 47, no. 13–15, pp. 2312–2327, Nov. 2020, doi: 10.1080/02664763.2019.1630372.

[6] B. Peng, R. K. Yu, K. L. DeHoff, and C. I. Amos, "Normalizing a large number of quantitative traits using empirical normal quantile transformation," *BMC Proc.*, vol. 1, no. S1, pp. 1–5, Dec. 2007, doi: 10.1186/1753-6561-1-S1-S156.

[7] G. D. Rayner and H. L. MacGillivray, "Weighted quantile-based estimation for a class of transformation distributions," *Comput. Stat. Data Anal.*, vol. 39, no. 4, pp. 401–433, Jun. 2002, doi: 10.1016/S0167-9473(01)00090-1.

[8] T. Zhang and B. Yang, "Box–Cox transformation in big data," *Technometrics*, vol. 59, no. 2, pp. 189–201, Apr. 2017, doi: 10.1080/00401706.2016.1156025.

[9] J. W. Osborne, "Improving your data transformations: Applying the Box–Cox transformation," *Pract. Assess. Res. Eval.*, vol. 15, no. 1, Jan. 2010, doi: 10.7275/QBPC-GK17.

[10] Fish Market. Accessed: Feb. 26, 2025. [Online]. Available:https://www.kaggle.com/datasets/vipullrathod/fish-market.

[11] *Models and test, we have used.* [Online]. Available: http://jse.amstat.org/datasets/fishcatch.txt.

[12] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Brief Bioinform.*, vol. 24, no. 2, pp. 1–11, Mar. 2023, doi: 10.1093/bib/bbad002.

[13] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, pp. 1–45, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[14] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," *arXiv preprint*, Oct. 2018. Accessed: Mar. 1, 2025.

[15] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 26, pp. 758–763, 2019, doi: 10.1007/978-3-030-03146-6_86.

[16] F. Kazemi, N. Asgarkhani, T. Shafighfard, R. Jankowski, and D. Y. Yoo, "Machine-learning methods for estimating performance of structural concrete members reinforced with fiber-reinforced polymers," *Arch. Comput. Methods Eng.*, vol. 32, no. 1, pp. 571–603, Jan. 2024, doi: 10.1007/s11831-024-10143-1.

[17] S. von Bülow, G. Tesei, and K. Lindorff-Larsen, "Machine learning methods to study sequence–ensemble–function relationships in disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 92, p. 103028, Jun. 2025, doi: 10.1016/j.sbi.2025.103028.

[18] B. Liu, Y. Yu, Z. Liu, Z. Cui, and W. Tian, "Prediction of CO_2 solubility in aqueous amine solutions using machine learning method," *Sep. Purif. Technol.*, vol. 354, p. 129306, Feb. 2025, doi: 10.1016/j.seppur.2024.129306.

[19] P. D'Orazio and A. D. Pham, "Evaluating climate-related financial policies' impact on decarbonization with machine learning methods," *Sci. Rep.*, vol. 15, no. 1, p. 1694, Dec. 2025, doi: 10.1038/s41598-025-85127-7.

[20] N. Tuerxun *et al*., "Accurate estimation of jujube leaf chlorophyll content using optimized spectral indices and machine learning methods integrating geospatial information," *Ecol. Inform.*, vol. 85, p. 102980, Mar. 2025, doi: 10.1016/j.ecoinf.2024.102980.

[21] K. Bogner, F. Pappenberger, and H. L. Cloke, "Technical note: The normal quantile transformation and its application in a flood forecasting system," *Hydrol. Earth Syst. Sci.*, vol. 16, no. 4, pp. 1085–1094, 2012, doi: 10.5194/hess-16-1085-2012.

[22] M. Buchinsky, "Quantile regression, Box–Cox transformation model, and the U.S. wage structure, 1963–1987," *J. Econom.*, vol. 65, no. 1, pp. 109–154, Jan. 1995, doi: 10.1016/0304-4076(94)01599-U.

[23] X. Xie, Y. Mei, B. Gu, and W. He, "Changing Box–Cox transformation parameter as an early warning signal for abrupt climate change," *Clim. Dyn.*, vol. 60, no. 11–12, pp. 4133–4143, Jun. 2023, doi: 10.1007/s00382-022-06563-z.

[24] J. Nagendra *et al*., "Evaluation of surface roughness of novel Al-based MMCs using Box–Cox transformation," *Int. J. Interact. Des. Manuf.*, vol. 18, no. 5, pp. 3369–3382, Jul. 2024, doi: 10.1007/s12008-023-01561-9.

[25] A. A. Al Abbasi, M. J. Alam, S. Saha, I. A. Begum, and M. F. Rola-Rubzen, "Impact of rural transformation on rural income and poverty for sustainable development in Bangladesh: A moments-quantile regression with fixed-effects models approach," *Sustain. Dev.*, vol. 33, no. 2, pp. 2951–2974, Apr. 2024, doi: 10.1002/sd.3276.

[26] Dhiman, H. S., Deb, D., & Guerrero, J. M. (2019). Hybrid machine intelligent SVR variants for wind forecasting and ramp events. *Renewable and Sustainable Energy Reviews*, *108*, pp. 369-379.

[27] Terrault, N. A., & Hassanein, T. I. (2016). Management of the patient with SVR. *Journal of hepatology*, *65*(1), pp. 120-129.

[28] Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PloS one*, *18*(2), e0279774.

[29] Gao, J. (2024). R-Squared (R2)–How much variation is explained?. *Research Methods in Medicine & Health Sciences*, *5*(4), 104-109.

[30] Tengtrairat, N., Woo, W. L., Parathai, P., Rinchumphu, D., & Chaichana, C. (2022). Non-intrusive fish weight estimation in turbid water using deep learning and regression models. *Sensors*, *22*(14), 5161.

[31] Hamzaoui, M., Aoueileyine, M. O. E., Romdhani, L., & Bouallegue, R. (2023). Optimizing XGBoost performance for fish weight prediction through parameter pre-selection. *Fishes*, *8*(10), 505.

[32] Mots' oehli, M., Nikolaev, A., IGede, W. B., Lynham, J., Mous, P. J., & Sadowski, P. (2024, July). Fishnet: Deep neural networks for low-cost fish stock estimation. In *2024 IEEE International Conference on Omni-layer Intelligent Systems (COINS)* (pp. 1-7). IEEE.

[33] Zhang, T., Yang, Y., Liu, Y., Liu, C., Zhao, R., Li, D., & Shi, C. (2024). Fully automatic system for fish biomass estimation based on deep neural network. *Ecological Informatics*, *79*, 102399.

[34] Rani, S. J., Ioannou, I., Swetha, R., Lakshmi, R. D., & Vassiliou, V. (2024). A novel automated approach for fish biomass estimation in turbid environments through deep learning, object detection, and regression. *Ecological Informatics*, *81*, 102663.