Oyster Meat Yield Estimation via Multimodal Fusion of Shape and Appearance Features with ViT and VAE

Zhipeng Liang
Yantai Huadong Electronics Technology Co., Ltd, Yantai, China

Xinqi Fu
Yantai Huadong Electronics Technology Co., Ltd, Yantai, China

Haijin Fu
Yantai Huadong Electronics Technology Co., Ltd, Yantai, China

JunFeng Zhang
Yantai Huadong Electronics Technology Co., Ltd, Yantai, China

Feng Zhao
Information Engineering College, Yantai Institute of Technology, Yantai China; School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China

Jinyu Hao
School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China

Yali Li
Information Engineering College, Yantai Institute of Technology, Yantai China

ABSTRACT
As an economically important species in aquaculture, quality classification and meat yield assessment of oysters are crucial for industrial efficiency. Traditional manual assessment methods are inefficient and subjective. While computer vision-based approaches have been explored for oyster weight estimation, they primarily rely on manually measured morphological parameters and often overlook valuable visual appearance features inherent in the raw images. Furthermore, weight alone is an insufficient indicator of meat content, as large shells may contain little meat. To address these limitations, this study pioneers a multimodal oyster meat yield prediction model that synergistically combines shape and appearance features for quality grading. Specifically, a segmentation network extracts shape parameters and appearance image data, constructing a multimodal dataset. A dual-branch feature extraction architecture is designed: the appearance branch utilizes self-attention mechanisms to capture pixel-level interactions, while the shape branch employs variational autoencoders (VAE) to map features into robust latent representations. These modality-specific features are concatenated and processed through a Multilayer Perceptron (MLP) to directly predict meat yield. Experimental results demonstrate that the proposed multimodal fusion approach, which comprehensively leverages both morphological and visual characteristics, establishes significantly more robust and accurate mapping relationships than unimodal models relying solely on shape or appearance. The model effectively captures complementary information and adaptively modulates cross-modal influences, thereby enhancing prediction accuracy (R^2=0.9567). The key advantages of the proposed method lie in its ability to overcome the limitations of manual feature measurement and unimodal analysis by automatically extracting and fusing richer information and achieve superior prediction performance crucial for practical quality grading applications in oyster aquaculture.

Keywords
Oyster Meat Yield Estimation, Multimodal Fusion Learning, Variational Autoencoder, Vision Transformer

1. INTRODUCTION
Oysters are becoming an important source of nutrition for humans due to their high protein and low fat properties [1]. With population growth and natural ecosystem degradation, the majority of oysters consumed daily originate from aquaculture, positioning them as an increasingly vital cash crop in this industry [2], [3]. In aquaculture, oyster value is predominantly determined by quality, which is primarily reflected through meat yield rate [4]. Consequently, accurate classification of oyster quality based on meat yield rate proves essential for aquaculture enterprises to maximize economic benefits [5]-[7]. However, traditional quality assessment methods relying on manual weighing or empirical judgments exhibit limitations including low sorting efficiency, labor intensiveness, and insufficient accuracy [8]-[11]. This underscores the urgent need for a simple yet effective technical approach to estimate oyster meat yield through external characteristics [12], enabling precise quality classification. In recent years, computer vision which is a non-invasive technique with great potential for meat yield estimation has received extensive attention from the academic community [13]-[18].

However, current domestic and international research predominantly focuses on weight estimation through morphological analysis to achieve selective breeding or quality classification [19]-[23]. These computer vision-based approaches establish relationships between weight and morphological features. Weight estimation methodologies can be categorized as single-factor or multi-factor based on

influential parameters [24], [25]. Single-factor methods investigate correlations between weight and individual morphological characteristics (e.g., length, width), constructing univariate regression models. For instance, Lim et al. demonstrated that length-weight relationships (LWR) serve as common tools for assessing the overall health status of cultured aquatic organisms [26]. Singh, Y. T. revealed shell weight showed positive correlation with length, breadth, width and dry meat weight, and abiotic parameters, silt and clay [27]. Single-factor weight estimation methods are simple and easy to implement [28], [29], but in actual culture, oysters of the same length or width may have some differences in morphology. Therefore, this type of method only considers the relationship between individual morphological characteristics and weight, and has some limitations when classifying quality based on weight estimation.

Multi-factor weight estimation methods explore multivariate relationships between weight and combined morphological features (e.g., area, perimeter), establishing comprehensive predictive models [30], [31]. Dame et al. conducted allometric analyses of shell weight, total weight, dry/wet meat weight, height, and length combinations in oysters from South Carolina subtidal and intertidal zones, identifying height as the most effective predictor for biomass parameters [32]. Pineda‐Metz et al. employed length-width-height variables in random forest models to estimate total weight, shell weight, and soft tissue wet weight, demonstrating superior performance compared to allometric models [33]. Gimin, R. et al. confirmed strong correlations between live weight and shell dimensions (length, height, width) as well as shell volume [34]. Although computer vision-based weight estimation has achieved maturity in quality classification, inherent limitations persist. Specifically, oysters may exhibit substantial total weight with disproportionately high shell-to-meat ratios (i.e., large but hollow shells). Therefore, meat yield rate emerges as a more market-relevant quality indicator, emphasizing superior meat content per unit weight.

Studies have demonstrated significant correlations between oyster meat yield and morphological characteristics. For instance, Vu, S. V., et al. investigated the relationship between shell morphometric traits (cup ratio and fan ratio) and meat production (soft tissue weight and condition index) in Portuguese oysters, revealing genetic correlations between these shape ratios and meat output, thereby indicating the critical role of shell morphology in determining flesh productivity [35], [36]. Singh, Y. T. established quantitative relationships between length and total weight/shell weight/wet meat weight/dry meat weight, further explored length-meat yield correlations, and identified seasonal variations in condition index and meat production [8]. These findings confirm the feasibility of estimating meat yield through morphological analysis for quality classification. However, current research on oyster quality grading—whether based on weight or meat yield estimation—predominantly relies on manually measured shape parameters while neglecting inherent visual characteristics (e.g., color, texture) captured in raw imagery. This underscores the necessity to develop an automated multimodal approach integrating both appearance and morphological features for accurate meat yield estimation and subsequent quality classification.

To address the critical limitations of reliance on manual measurements, neglect of rich visual appearance cues, and the inherent insufficiency of weight as a sole indicator of meat content, this study proposes a novel and robust multimodal oyster meat yield prediction model. This model fundamentally shifts the paradigm by synergistically combining both shape and appearance features derived directly and automatically from images for precise quality grading. Crucially departing from prior unimodal or manually-feature-dependent approaches, the proposed method employs a segmentation network as the foundational step to extract both precise shape parameters and the corresponding appearance image data (focusing solely on the oyster), thereby constructing a truly integrated multimodal dataset that inherently captures potential plumpness-related features visible in the oyster's appearance. The main contribution of this work is twofold: (1) Pioneering multimodal fusion in bivalve meat yield prediction: This study introduces the first framework that jointly models morphological parameters and visual appearance features, enabling the capture of cross-dimensional interaction patterns typically overlooked by unimodal approaches. (2) Specialized dual-branch architecture: The designed network implements functional specialization through parallel processing streams. The self-attention branch enhances discriminative feature representation via pixel-wise correlation weighting, while the auto-encoder branch constructs probabilistic distributions of shape parameters through latent space modeling. Their synergistic operation significantly improves predictive performance.

2. MATERIAL

In the experiment, 184 oyster samples were collected from aquaculture farms in Shandong Province, China, using a strict random sampling method to ensure randomness and representativeness. Images were captured using a Canon EOS 5D Mark IV camera (Canon Inc., China), and weighing was performed with a JA3003 electronic analytical balance (sensitivity 1 mg, Shanghai Precision & Scientific Instrument Co., Ltd., China) to measure both the wet weight of the soft tissue and the total wet weight, which were used to calculate the meat yield ratio (meat yield ratio = wet weight of soft tissue / total wet weight). Additionally, a S102-107-101 vernier caliper (Shanghai Yonghui Industrial Development Co., Ltd., China) was used to manually measure length, width, and other shape features to validate the accuracy of the shape features measured by the machine learning method. The oysters were placed on a black background, with the camera fixed above, maintaining a consistent distance from the surface, and each oyster was photographed from both the front and back. The camera settings included an exposure time of 1/80 s, ISO 12800, a focal length of 100 mm, and an image resolution of 4480 × 4480 pixels. This meticulous data collection process ensures the accuracy and reliability of the data, crucial for validating the effectiveness of the proposed multimodal oyster meat yield prediction model. Samples were collected as shown in Fig.1.

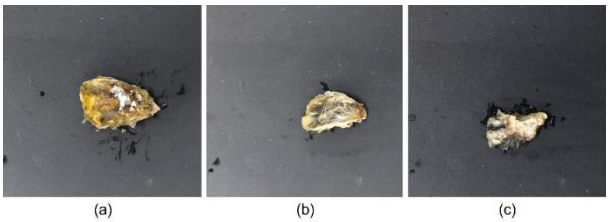

Fig 1: Samples diagrams for various forms of oysters

In this experiment, the collected raw images were used to create a multimodal dataset for training a multimodal model. This dataset includes both apparent image data and numerical shape data.

The process of creating the apparent image dataset involved first collecting raw image data containing the target (e.g., oysters) and preprocessing these images by resizing and normalization. Next, a UNet-style network model was used to segment the target regions. By training the UNet model to generate segmentation masks and applying these masks to the original images, the background areas were blacked out, leaving only the target regions, thus generating the apparent data, as shown in Fig.2.

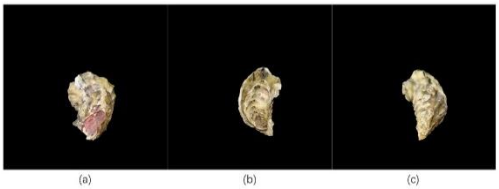

Fig 2: Samples of apparent image dataset

The process of making the shape feature dataset is as follows: the contour of the target area obtained by image segmentation is used to extract the shape feature numerical data using a series of digital image processing methods. Specifically, the segmentation masks were used to extract the target's contours and measure various shape features, such as Length, Width, Area, Perimeter, Convex Hull Length, and Convex Hull Area. Subsequently, the shape data were standardized to ensure that the shape feature values were independent of the oyster's size, resulting in new shape feature attributes: Length Eccentricity, Width Eccentricity, Roughness, Compactness, Elongation, and Plumpness. Detailed descriptions of these attributes are provided in Table 1. The attributes marked with an underline are the normalized attributes, which are the actual attributes used for model training.

Table 1. Sample shape feature attribute description

Feature attributes	Description of feature attributes
Length	The distance between the two furthest points in the contour.
Width	The distance between the two furthest points in the contour perpendicular to Length.
Area	The pixel area of oyster.
Perimeter	The length of oyster contour.
Convex Hull Length	The length of the convex hull of the set of **Perimeter** points.
Convex Hull Area	The pixel area of convex hull.
Length Eccentricity	According to the intersection point of **Length** and **Width**, the **Length** is divided into two segments, **Length Eccentricity** means the ratio of the short segment to the long segment.
Width Eccentricity	The ratio of the short segment to the long segment, which is divided by **Width**.
Roughness	The **Perimeter** verses **Convex Hull Length** ratio.
Compactness	This is calculated as $p^2/(4\pi A)$,where p is the **Perimeter**, A is the **Area**.
Elongation	The ratio of **Length** to **Width**.
Plumpness	The ratio of **Area** to **Convex Hull Area**.

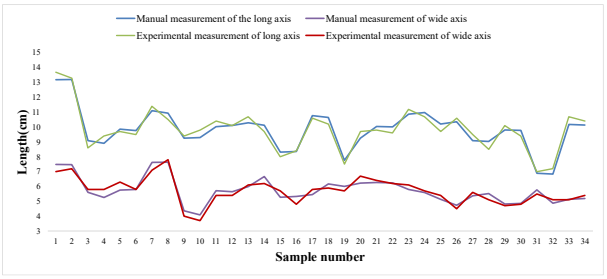

Fig 3: Comparison of experimental and manual measurement data

In order to verify the validity of the experiment, the experimentally measured long and width axes were compared with the manually measured long and short axes, and the experimental results are shown in Fig.3, where the experiments of 34 samples are shown due to the limited space. The green and purple color in the figure are the experimentally measured long and short axes of oysters, and the blue and red color are the manually measured long and short axes of oysters. It can be seen from the figure that the experimentally measured data and the manually measured data are similar, and the data obtained by the two measurement methods are highly consistent, which verifies the effectiveness of the proposed feature extraction method. This part is described in detail in the previous work [37].

Finally, the apparent data and shape features were integrated into a multimodal dataset for training the multimodal model. By doing so, the model can simultaneously utilize the apparent image information and shape features, thereby enhancing the training effectiveness and improving the model's performance.

3. METHOD

3.1 Multimodal fusion learning model construction

Relevant studies have shown that there is a significant correlation between the meat yield of oysters and its features such as shape, texture and color. In this paper, a prediction model of oyster meat yield based on multimodal fusion learning is proposed. The model constructs a complete multimodal fusion learning framework containing shape feature extraction network, apparent feature extraction network, feature fusion module and regression prediction by deeply fusing the shape features and apparent features of oysters, as shown in Fig.4.

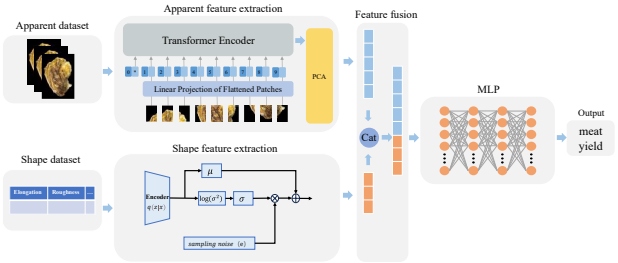

Fig 4: Overall structure of the multimodal feature fusion learning model

3.2 Apparent feature extraction based on self-attention mechanism

This section details the appearance feature extraction method based on the Vision Transformer (ViT) [38] and Principal Component Analysis (PCA). First, the ViT encoder is utilized to extract global image features, followed by dimensionality

reduction via PCA to address potential overfitting issues arising from high-dimensional features when the feature dimensionality significantly exceeds the sample size. ViT overcomes the limitations of convolutional neural networks (CNNs) in handling long-range dependencies by prepending a class token to the sequence and employing linear layers for classification. Leveraging self-attention mechanisms, ViT captures long-range dependencies adaptively without manual design of convolutional kernels, enabling autonomous learning of interactive features. This flexibility ensures superior performance in complex pattern recognition tasks, justifying its selection for oyster appearance feature extraction. The core components of ViT in the model include image patching, class token insertion, positional encoding, and the encoder architecture, as illustrated in Fig.5.

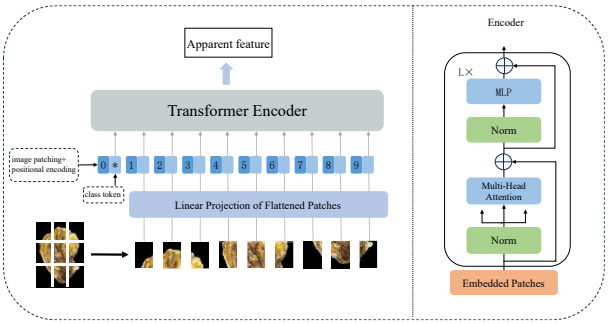

Fig 5: Structure of the apparent feature extraction model

Specifically, the image is divided into 16×16 patches, resulting in 196 patches for a 224×224 image. Each patch is flattened and linearly embedded into a 768-dimensional vector, forming a 2D matrix of size [196,768]. A special class token (clsToken), serving as a trainable parameter with dimensions [1,768], is concatenated to the front of the patch vectors, producing an augmented matrix of size [197,768]. Positional encoding, implemented as a trainable parameter, is added element-wise to the matrix without altering its dimensions, maintaining the shape [197,768]. This matrix is then fed into the encoder for feature extraction, with the output retaining dimensions [197,768]. The first vector (corresponding to the clsToken) from this output is extracted as the global feature representation of the entire image, yielding a [1,768] feature matrix for subsequent analysis.

The Multi-Head Self-Attention (MHSA) mechanism is one of the core components of the Transformer architecture. It enables the model to capture the relationships and features of many different aspects of the input data by processing multiple different self-attention heads in parallel. The model processes the input sequence through a self-attention mechanism and a feed-forward network that progressively extracts and integrates image features through multiple such layers. Residual connectivity and layer normalization ensure the flow and stability of information across the layers. The self-attention mechanism is used to process the input sequence Z, where Z contains the information extracted from the image. The self-attention mechanism allows the model to process the information at each position while being able to attend to the information at other positions in the sequence. This is accomplished by calculating the attention weights with the following formula:

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where Q、K、V denote the linear mapping of query, key, and value, respectively, and d_k is the mapping dimension. In practice, in order to enhance the expressive power of the model, it is common to use the multi-head self-attention mechanism, i.e., to compute multiple sets of Q、K、V, and then stitch the results together.

$$MultiHeadSelfAttention(Z) = Concat(head_1, head_2, \cdots, head_h)W_O \quad (2)$$

Here $head_i = Attention(ZW_{Qi}, ZW_{Ki}, ZW_{Ki})$, W_O is the final output mapping weight. The output of the self-attention mechanism is further processed through the feedforward network. The feedforward network consists of two fully connected layers with an activation function (usually ReLU) in the middle. This helps the model learn more complex features and patterns.

$$Y' = FeedForward(Y) = \mathrm{ReLU}(TW_{F1} + b_{F1})W_{F2} + b_{F2} \quad (3)$$

Where Y is the output of the self-attention mechanism. To maintain the flow of information and the stability of the model, residual connectivity and layer normalization are introduced. The residual connection passes information directly to the next layer by adding the input sequence Z to the output of the feedforward network Y'.

$$Y'' = LayerNorm(Y + Z) \quad (4)$$

Here, it is the $LayerNorm$ operation that helps mitigate the problem of vanishing gradients during training and improves the stability of the model during training. This structure allows the model to better capture the information in the input sequence while retaining important features of the original input. The multi-head self-attention mechanism is versatile, i.e., different attention heads can capture different aspects of the relationships and features of the input data, thus enhancing the expressive power of the model; at the same time, by computing multiple attention heads in parallel, the parallel computing capability of modern hardware can be effectively utilized to improve the computational efficiency; in addition, the scaled dot product attention mechanism stabilizes the gradient and avoids the gradient explosion during training problem.

However, when the feature dimension (768) is much larger than the number of samples (368), the data is extremely sparse in the high-dimensional space, which tends to cause the model to memorize the noise in the training data instead of learning the generalization laws, and also requires more arithmetic resources.

Therefore, PCA is chosen to perform dimensionality reduction on the extracted global features, and the feature directions that contribute the most to the data distribution are filtered out by retaining the principal components with the largest variance. Therefore, we choose PCA to perform dimensionality reduction on the extracted global features, and filter out the feature directions that contribute the most to the data distribution by retaining the principal components with the largest variance. However, traditional PCA relies on the eigenvalue decomposition of the covariance matrix, which requires sufficient samples (N > 768) to avoid the instability of the decomposition caused by the singularity of the covariance matrix, and since our sample size is insufficient, we use the singular value decomposition (SVD) to directly deal with the original matrix. However, traditional PCA relies on the eigenvalue decomposition of the covariance matrix, which requires sufficient samples (N > 768) to avoid the instability of

the decomposition caused by the singularity of the covariance matrix, and since the sample size is insufficient, singular value decomposition (SVD) is used to directly deal with the original matrix.

Firstly, the raw data need to be centered, i.e., by the following equation:

$$X_{centered} = X - \mu \tag{5}$$

Normalize the mean of each feature dimension to zero, where μ is the mean vector of each feature dimension. The purpose of this step is to remove the data bias and ensure the stability of the subsequent analysis. The data matrix $X_{centered} \in \mathbb{R}^{368\times768}$ obtained after centering contains 768 dimensional features for 368 samples.

Next, this matrix is decomposed by singular value decomposition (SVD) with the mathematical expression:

$$X_{centered} = U\Sigma V^T \tag{6}$$

where $U \in \mathbb{R}^{368\times368}$ is the left singular vector matrix, $\Sigma \in \mathbb{R}^{368\times768}$ is the diagonally expanded matrix containing the singular values, and $\Sigma \in \mathbb{R}^{768\times768}$ is the right singular vector matrix, whose column vectors represent the directions of the principal components in the original feature space. The variance contribution ratio is calculated based on the singular values with the formula:

$$CRV_i = \frac{\sigma_i}{\sum_{j=1}^{\min(N-1,d)}\sigma_j} \tag{7}$$

where σ_i is the ith singular value, characterizing the strength of variance in the direction of the corresponding principal component. In order to reduce the dimensionality, the principal components corresponding to the first k largest singular values need to be selected, and here, to prevent overfitting, they are selected up to 64 dimensions and the cumulative variance is greater than 95%, i.e., the first 64 columns are extracted from the right singular matrix V to form the projection matrix $W \in \mathbb{R}^{768\times64}$.Finally, by multiplying the centered data matrix with the projection matrix:

$$X_{pca} = X_{centered} \cdot W \tag{8}$$

The original 768-dimensional features can be mapped into a 64-dimensional low-dimensional space to complete the data dimensionality reduction. This process effectively compresses the data dimensions by retaining the principal components with the largest variance, while preserving the original information to the greatest extent.

3.3 Shape Feature Extraction based on Variational Autoencoder

Variational Autoencoder (VAE) as a generative model is able to learn latent representations of data, which cannot be directly observed or measured in machine learning and statistics. Latent spaces are hidden, but they influence and determine the distribution and structure of the data that can be observed, and can help interpret and capture the intrinsic structure and characteristics of the data. Compared to traditional auto-encoders, VAEs not only learn a compact representation of the data, but also automatically regularize VAEs during the training process due to the introduction of KL dispersion, which helps to prevent overfitting and improves the generalization ability of the model. The core advantage of VAEs is that they are able to capture complex nonlinear relationships in the data and map high-dimensional data to a structured latent space well. This mapping not only preserves the key features of the data, but also reflects the underlying structure and distribution of the data. Since the latent space of VAE is characterized by continuity and structure, and is made close to the standard normal distribution by the KL dispersion constraint, each dimension in the latent variables may correspond to some abstract features in the data, which may be decoupled and beneficial to the downstream tasks. Based on the above characteristics, the encoder part of the VAE is selected for oyster shape feature extraction in this study.

The structure of the shape feature extraction model is shown in Fig.6. The encoder converts the input data x into the mean μ and the logarithmic variance $log(\sigma)$ of the latent variable z. These parameters are used to describe the distribution in the latent space. Inputting the data into the encoder, the encoder outputs two codes: one is the original code $\mu(\mu_1、\mu_2、\mu_3)$; and the other is the code $\sigma(\sigma_1、\sigma_2、\sigma_3)$ of the control noise, which serves to assign weights to the random noise $e(e_1、e_2、e_3)$. Finally, the original coding and the weighted noise coding are summed to obtain the output of the VAE in the encoder part-the latent vector $z(z_1、z_2、z_3)$.

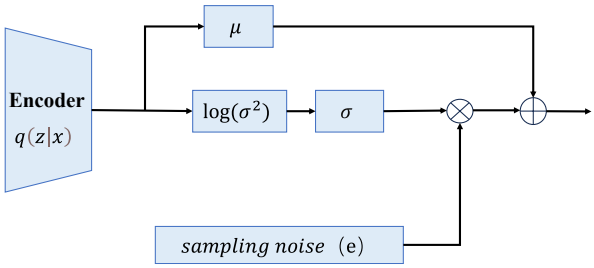

Fig 6. Structure of the data feature extraction model

Table 2. Evaluation indexes of regression results

Feature Vectors	RMSE	MAE	R^2
μ	0.0321	0.0296	0.9118
z	**0.0256**	**0.0277**	**0.9224**

The loss function of the VAE consists of two components: Reconstruction Error and KL Divergence. The reconstruction error measures the discrepancy between the original data and the reconstructed data, evaluating the model's ability to reconstruct the input x given the latent variable z. The formula is as follows, where $q(x|z)$ denotes the latent distribution generated by the encoder:

$$ReconstructionError = -E_{q(z|x)}[log\ p\ (x|z)] \tag{9}$$

The KL Divergence quantifies the difference between the latent distribution $q(x|z)$ output by the encoder and the standard normal distribution $p(z)$. The formula is:

$$KL\ Divergence = D_{KL}\big(q(z|x)\|p(z)\big) = -\frac{1}{2}\sum_{i=1}^{K}\big(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2\big) \tag{10}$$

Here, $q(x|z)$ is the posterior distribution of z given x, and $p(z)$ is the prior distribution of z. The complete loss function is:

$$L = D_{KL}\big(q(z|x)\|p(z)\big) - E_{q(z|x)}[\log p(x|z)] \tag{11}$$

In this study, the input raw data has a shape of [368, 6], and the latent dimension d is selected. During training, a complete model was constructed to evaluate the effectiveness of feature extraction. This includes an encoder with hidden layers implemented using fully connected layers (Dense Layers), and a decoder symmetrically structured to the encoder. Notably,

ReLU activation was used to enhance nonlinearity, while Sigmoid output ensured compatibility with normalized data.

To determine the optimal dimensionality of the latent space, experimental validation was conducted. Given the limited sample size, excessively high latent dimensions could lead to redundancy or ineffective encoding of useful information. Therefore, the latent dimension was constrained to a reasonable range. Experiments were performed with latent dimensions $d \in \{4,5,6,7\}$, and model performance was assessed based on reconstruction error and the predictive accuracy of shape features. The results, as shown in Table 3, indicate that when $d = 6$, the model achieved the best balance between reconstruction capability and generalizability. Consequently, a 6-dimensional vector was selected as the final output for shape feature extraction.

Table 3. Model performance under different potential dimensions

Latent Dimension	Reconstruction Error	R^2
4	0.12	0.85
5	0.09	0.88
6	0.08	0.92
7	0.07	0.87

3.4 Feature Vector Fusion and Regression Prediction

The This study developed an oyster meat yield estimation model based on regression prediction algorithms, aiming to achieve accurate predictions. By fusing feature vectors from multiple models, the overall predictive capability and robustness were significantly enhanced. Experimental results demonstrated that the concatenation (Concat) method for feature vector fusion, followed by input into a MLP for regression, outperformed the Add method, as detailed in Table 4.Using Concat for feature fusion improved the regression model's R^2 by 0.033, reduced MAE by 0.0044, and lowered RMSE by 0.0031, highlighting its substantial advantage in enhancing model performance.

Table 4: Experimental results comparison of feature fusion methods

Method	RMSE	MAE	R^2
Concat	**0.0193**	**0.0236**	**0.9567**
Add	0.0224	0.028	0.9237

In the feature fusion process, the appearance feature vector with 768 dimensions extracted by the multi-head self-attention mechanism is first reduced to 64 dimensions through Principal Component Analysis, while retaining the shape feature vector with 6 dimensions extracted by VAE. Subsequently, the Concat method is used to concatenate them, generating a 70-dimensional feature vector. This feature vector serves as input to the MLP regression predictor for oyster meat yield prediction. To construct a reasonable MLP regression prediction model, the following network structure is designed based on the 70-dimensional input features: the input layer contains 70 neurons, matching the dimension of the concatenated feature vector; the first hidden layer contains 128 neurons, approximately twice the input dimension, which expands feature expression capability while avoiding parameter explosion, with ReLU activation function; the second hidden layer contains 64 neurons, halving layer by layer to gradually compress redundant information while retaining key features; the third layer is further compressed to 32 dimensions to reduce model complexity and prevent overfitting; the output layer has 1 dimension for direct meat yield prediction. Batch normalization is added after each hidden layer to stabilize the training process and improve generalization capability. Meanwhile, Dropout is set to 0.2 for the first two hidden layers and 0.1 for the final layer, suppressing noise in early stages while preserving effective features in later stages. Through multiple experimental validations, this three-hidden-layer design can effectively capture nonlinear relationships in input features while avoiding training difficulties caused by excessive network depth.

To verify the rationality of the constructed meat yield prediction model, the model's prediction results are compared with manual measurement results. As shown in Fig 7, the blue curve represents the actual meat yield from manual measurements, while the orange curve represents the model's prediction results. It can be observed that the prediction results largely match the real data, indicating that the multimodal model constructed in this study has high feasibility and accuracy in oyster meat yield prediction.

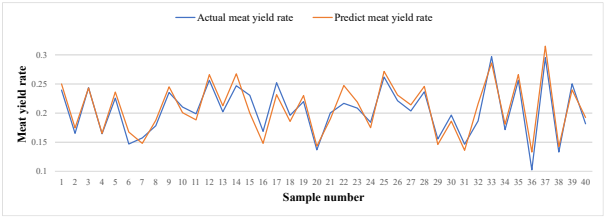

Fig 7. Comparison between model predictions and manual measurements of meat yield

4. Experiments and Results

4.1 Experimental Design

This section details the specific implementation and comparative experimental results of oyster meat yield estimation using the constructed multimodal dataset and the proposed multimodal prediction model. To demonstrate the effectiveness of the proposed method, this study designed three sets of experiments: (1) To validate the accuracy of the end-to-end model, i.e., to verify that direct meat yield prediction using the multimodal dataset yields more accurate results. The indirect method (predicting soft body wet weight and total weight separately before calculating meat yield) was compared with the direct meat yield prediction approach. (2) To verify the effectiveness of the multimodal oyster meat yield estimation strategy combining appearance and shape features, the proposed model was compared with unimodal regression prediction methods using either appearance features or shape features alone. (3) To validate the outstanding performance of the constructed multimodal prediction model in both feature extraction and regression prediction components, the feature extraction module, feature fusion module, and regression prediction module were compared with other commonly used models.

The computer configuration used in experiments was as follows: CPU frequency of 4.89GHz, 16GB memory, Windows11 (64-bit) operating system. The programming language was python3.8, with anaconda3 as the integrated development environment, and experiments were conducted using the pytorch framework. The Adam optimizer was employed.

4.2 Evaluation Metrics

The experimental results of regression prediction were evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2), calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \tilde{y}_i| \quad (12)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2 \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2} \quad (14)$$

$$R^2 = \frac{\left(\sum_{i=1}^{N}(y_i-\bar{y}_i)(\tilde{y}_i-\hat{y}_i)\right)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2\sum_{i=1}^{N}(\tilde{y}_i-\hat{y}_i)^2} \quad (15)$$

Where y_i is the true value, $\tilde{y}_i$ is the predicted value, $\bar{y}_i$ is the mean of true values, $\hat{y}_i$ is the mean of predicted values, and N is the number of predicted values.

4.3 Comparative Analysis of Direct and Indirect Prediction Results

Numerous studies have demonstrated correlations between oyster shape characteristics and both soft body wet weight and total weight. Given that meat yield = soft body wet weight / total wet weight, meat yield can be measured through two approaches: one involves separately predicting soft body wet weight and total weight before calculating meat yield, while the other directly predicts meat yield. To validate the superiority of the direct prediction method, this study compared the performance differences between the indirect calculation method (deriving meat yield from predicted soft body wet weight and total weight) and the direct meat yield prediction method. The experimental results are shown in Table 5. It should be noted that all models used the same shape features.

Table 5. Comparison of direct and indirect meat yield prediction results

	RMSE	MAE	R^2
Total Weight	0.1733	0.1358	0.2450
Soft Body Wet Weight	0.1321	0.1245	0.8808
indirect meat yield prediction	0.1452	0.1385	0.5629
direct meat yield prediction	0.0256	0.0277	0.9224

The experimental results indicate significant error accumulation effects in the indirect prediction method. Specifically, since meat yield is a ratio of two predicted values, errors propagate through the division operation. The error propagation formula can estimate meat yield errors as follows:

$$\frac{\Delta Y}{Y} \approx \left|\frac{\Delta M}{M}\right| + \left|\frac{\Delta W}{W}\right| \quad (16)$$

Where ΔM and ΔW represent prediction errors for soft body wet weight and total weight respectively, while M and W denote their predicted values. The RMSE of total weight prediction (0.1733) and MAE (0.1358) significantly amplify errors when propagated to meat yield calculation, whereas the direct method achieves a substantially lower RMSE of 0.0256. The direct prediction method outputs meat yield in one step through the MLP model, avoiding error accumulation from multi-step predictions. Furthermore, the R^2 of total weight prediction (0.2450) is significantly lower than that of soft body wet weight (0.8808), indicating shape features have weaker explanatory power for total weight. This may be because total weight is more influenced by non-morphological factors like shell weight, while soft body wet weight shows more direct correlations with morphological features (e.g., eccentricity). The R^2 of indirect meat yield prediction is markedly lower than the direct method, demonstrating how error propagation substantially reduces model explanatory power. Moreover, the direct method not only achieves lower errors but also simplifies the process from two-step prediction (total weight and soft body wet weight) to single-step output, reducing computational complexity. For scenarios requiring rapid meat yield estimation, the direct method offers advantages in real-time performance and resource efficiency. Additionally, it eliminates systematic biases caused by inaccurate total weight prediction, thereby enhancing model robustness.

In conclusion, the direct prediction method demonstrates significant advantages over the indirect method in error control, explanatory power, and practical utility, providing a reliable solution for efficient oyster meat yield estimation and validating the rationality of the proposed direct prediction approach.

4.4 Comparative Analysis of Multimodal Model and Unimodal Prediction Results

To validate the effectiveness of the multimodal meat yield prediction strategy combining oyster appearance features and shape features, this study designed comparative experiments to evaluate the performance between the proposed multimodal model and unimodal prediction strategies using either appearance features or shape features alone. The strategy workflow is illustrated in Fig 8.

The unimodal prediction models were divided into two categories: one using only appearance image information, consisting of an appearance feature extraction network and MLP, where the feature vector obtained through ViT-PCA feature extraction was directly input into the MLP for regression prediction; the other using only shape feature information, consisting of a shape feature extraction network and MLP, where only the feature vector output from the VAE encoding part was input into the MLP for regression prediction. The multimodal model concatenated the feature vector obtained from the ViT-PCA feature extraction network for appearance image information with the feature vector obtained from the VAE encoding feature extraction network for shape information, and then input the combined vector into the MLP model for regression prediction. Additionally, to verify the necessity of the feature extraction module, unimodal models with and without feature extraction were compared. The hyperparameters of the ViT-PCA, VAE, and MLP models remained consistent across all models. The prediction performance was evaluated using three metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

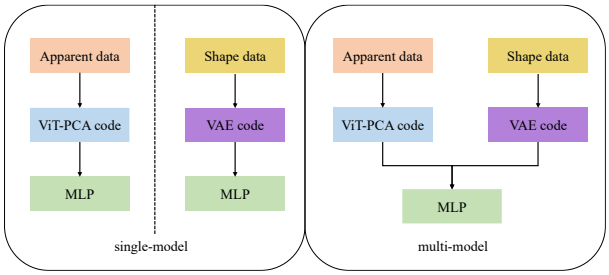

Fig 8. Schematic diagram of unimodal prediction & multimodal prediction process

The experimental results are shown in Table 6. To confirm the effectiveness of feature extraction, unimodal prediction results with and without feature extraction were compared. For shape information, model performance significantly improved after VAE encoding, with RMSE decreasing from 0.0442 to 0.0256 and R^2 increasing from 0.8805 to 0.9224. For appearance information, after ViT feature extraction, regression performance R^2 improved from 0.7651 to 0.8805. Most importantly, the multimodal model combining appearance features (ViT-PCA) and shape features (VAE encoding) achieved RMSE=0.0193 and R^2=0.9567, significantly outperforming unimodal models. These results demonstrate the notable complementarity between appearance and shape features, which, when integrated, can effectively enhance prediction accuracy.

To provide a more intuitive comparison, Fig 9 visualizes the prediction performance of the multimodal and unimodal models using scatter plots. In Fig 9(a), the predicted values generated by the multimodal model are tightly clustered along the ideal line (y = x), indicating high prediction precision and minimal deviation. The coefficient of determination reaches R^2 = 0.9561, confirming strong agreement between predicted and actual values. In contrast, Fig 9(b) illustrates the unimodal model based solely on appearance features, where the predictions exhibit more pronounced deviations, especially in higher yield ranges, resulting in a lower R^2 = 0.8422. These visual comparisons further validate the superiority of the multimodal strategy, highlighting its ability to leverage complementary information from both shape and appearance modalities to improve predictive performance.

Table 6. Comparison of multimodal model and unimodal model prediction results

Model	Input	Method	RMSE	MAE	R^2
Without Feature Extraction	Apparent	MLP	0.0464	0.0396	0.7651
	Shape	MLP	0.0453	0.0358	0.8744
With Feature Extraction	Apparent	Vit-PCA+MLP	0.0442	0.0348	0.8805
	Shape	VAE+MLP	0.0256	0.0277	0.9224
Multi-modal	**Apparent+Shape**	**VAE+ViT-PCA+MLP**	**0.0194**	**0.023**	**0.9567**

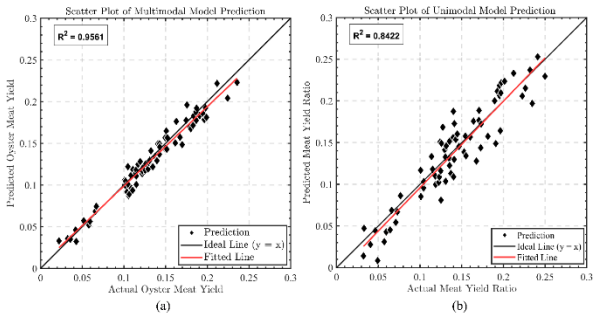

Fig 9: Comparison of prediction performance between multimodal and unimodal models based on scatter plots.

Additionally, as shown in Table 7, the rationality of appearance image data in multimodal dataset construction was validated. Experimental results demonstrate that performing feature extraction directly on original images yielded significantly worse regression prediction results for oyster yield compared to using segmented target regions. Although the original dataset used black backgrounds, the segmented data still showed clear differences in feature representation, further confirming the necessity of retaining only target regions in collected images during dataset preparation.

Table 7: Comparison of model regression effects before and after masking of the epistatic dataset

input	RMSE	MAE	R^2
Apparent data-original image	0.0455	0.0356	0.8786
Apparent data-Mask Image	**0.0442**	**0.0348**	**0.8805**

In summary, while unimodal information can provide basic yield predictions, the multimodal strategy combining appearance and shape information clearly outperforms unimodal approaches. This further confirms the complementary nature of different modal information, demonstrating that utilizing multimodal information can yield more accurate prediction results, validating the effectiveness of the proposed multimodal fusion learning method for oyster yield estimation.

4.5 Comparative Analysis of Various Multimodal Model Construction Results

The multimodal model consists of three core components: feature extraction, feature fusion, and regression prediction. Although each component has multiple excellent alternative methods, experimental verification shows the current model performs exceptionally well in oyster yield prediction. Specifically, ViT-PCA was used to extract appearance features (adjusted to 64 dimensions) and VAE for shape feature extraction. Finally, simple concatenation (concat) was used for feature fusion, with MLP performing regression prediction to achieve accurate oyster yield estimation.To validate the rationality of feature extraction in the model, ablation experiments were conducted. First, the appearance feature extractor was replaced with ResNet, a classical convolutional network whose residual structure effectively captures local detail features. Second, the shape feature extractor was substituted with an Autoencoder (AE). Each component was replaced separately or both were replaced simultaneously to comprehensively evaluate the feature extraction module's impact. To assess the regression model's rationality, MLP was compared with polynomial regression. Since previous experiments demonstrated concat's superior performance in feature fusion, this method was maintained to focus on analyzing the feature extraction module.

Table 8: Comparison of experimental results of multiple multimodal models

Model	RMSE	MAE	R^2
ResNet+AE+MLP	0.0918	0.0553	0.4844
ResNet+VAE+MLP	0.0403	0.0318	0.9010
ViT-PCA+AE+ polynomial	0.1581	0.1237	-1.2272
ViT-PCA+AE+MLP	0.0283	0.0216	0.9512
ViT-PCA+VAE+MLP	0.0073	0.0050	0.9967

As shown in Table 8, the constructed multimodal fusion learning model demonstrates optimal performance in regression prediction. When replacing both the appearance and shape feature extractors with ResNet and Autoencoder (AE) respectively, the model performance significantly deteriorates (R^2=0.4844), representing the most substantial performance decline among all replacement experiments. Analyzing each feature extraction module separately reveals that the choice of appearance feature extraction method substantially impacts model performance. Models employing self-attention mechanisms achieve R^2 values of 0.9312 and 0.9567, while those using ResNet yield R^2 values of 0.4844 and 0.9010. This discrepancy indicates that self-attention mechanisms more effectively capture global appearance features, whereas ResNet's local feature extraction characteristics may limit its modeling capability for complex oyster appearance patterns. Furthermore, in shape feature extraction methods, VAE significantly outperforms AE, with ViT-PCA+VAE+MLP (R^2 =0.9567) showing a 2.7% improvement over ViT-PCA+AE+MLP (R^2=0.9312), demonstrating that VAE's latent space regularization constraints enable better extraction of discriminative shape features. Comparing MLP with polynomial regression shows MLP's superior performance in modeling nonlinear relationships. For instance, Trans+AE+MLP achieves R^2 =0.9312, while Trans+AE+Polynomial yields R^2=-1.2272, performing worse than mean prediction. This suggests polynomial regression's insufficient model complexity fails to capture nonlinear relationships between oyster yield and multimodal features, whereas MLP's multilayer nonlinear transformations enable more precise fitting, further validating deep regression models' necessity for this task.

In conclusion, the multimodal model employing ViT-PCA for appearance feature extraction, VAE for shape feature extraction, concatenation for feature fusion, and MLP for regression prediction demonstrates outstanding performance in oyster yield prediction. This model effectively validates the multimodal fusion approach's efficacy.

5. CONCLUSIONS

This chapter proposes an oyster yield prediction method based on multimodal fusion learning, constructing an efficient multimodal prediction framework by comprehensively considering the influence of both shape and appearance features on yield. The method innovatively achieves deep integration of appearance and shape features through feature concatenation, significantly enhancing model prediction performance. To enable non-invasive oyster yield estimation, a multimodal dataset was constructed using collected images, where appearance image data highlights target regions through U-Net segmentation, and shape attribute data is measured based on obtained contours.Experimental results demonstrate the constructed multimodal oyster yield prediction model's excellent regression performance (R^2=0.9567), fully validating the proposed method's effectiveness. The study reveals that feature concatenation enables the model to effectively capture complementary information across modalities, achieving precise yield prediction. Appearance features significantly influence yield prediction, and their combination with shape features proves both necessary and effective for regression prediction.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

7. REFERENCES

[1] Asha, K. K., Anandan, R., Mathew, S., & Lakshmanan, P. T. (2014). Biochemical profile of oyster Crassostrea madrasensis and its nutritional attributes. The Egyptian Journal of Aquatic Research, 40(1), 35-41.

[2] Yearbook F. Fishery and aquaculture statistics 2016 [J]. FAO: Rome, Italy, 2019.

[3] Botta, R., Asche, F., Borsum, J. S., & Camp, E. V. (2020). A review of global oyster aquaculture production and consumption. Marine Policy, 117, 103952.

[4] Mizuta, D. D., & Wikfors, G. H. (2019). Seeking the perfect oyster shell: a brief review of current knowledge. Reviews in Aquaculture, 11(3), 586-602.

[5] Botta, R., Asche, F., Borsum, J. S., & Camp, E. V. (2020). A review of global oyster aquaculture production and consumption. Marine Policy, 117, 103952.

[6] Nayar, K. N., Mahadevan, S., & Muthiah, P. (1987). Economics of oyster culture. CMFRI Bulletin-Oyster culture-status and prospects, 38, 67-70.

[7] Van In, V., & O'Connor, W. (2024). Blue Economy: Valuing the Carbon Sequestration Potential in Oyster Aquaculture.

[8] Lapico A, Sankupellay M, Cianciullo L, et al. Using image processing to automatically measure pearl oyster size for selective breeding [C]. In 2019 Digital Image Computing: Techniques and Applications (DICTA), 2019: 1–8.

[9] VG N, Hareesh K. Quality inspection and grading of agricultural and food products by computer vision-a review [J]. International journal of computer applications, 2010, 975: 8887.

[10] Narendra, V.G., Hareesh, K.S., 2010. Quality inspection and grading of agricultural and food products by computer vision-a review. Int. J. Comput. Appl. 2 (1), 43–65.

[11] Ye, X., Liu, Y., Zhang, D., Hu, X., He, Z., Chen, Y., 2023. Rapid and accurate crayfish sorting by size and maturity based on improved YOLOv5. Appl. Sci. 13 (15), 8619

[12] Yuan, B., Cui, Y., Liu, W., Sheng, W., Xu, H., & Yang, L. (2023). Consumer preferences for oyster trait attributes in China: A choice experiment. Aquaculture, 571, 739471.

[13] Parr, M. B., Byler, R. K., Diehl, K. C., & Hackley, C. R. (1995). Machine vision based oyster meat grading and sorting machine. Journal of Aquatic Food Product Technology, 3(4), 5-24.

[14] Lee, D. J., Xu, X., Lane, R. M., & Zhan, P. (2004, December). Shape analysis for an automatic oyster grading system. In Two-and Three-Dimensional Vision Systems for Inspection, Control, and Metrology II (Vol. 5606, pp. 27-36). SPIE.

[15] Zhang L, Du X, Guo S. Sse-ppl: a machine vision technology to detect aquatic product quality [C].In International Conference on Physics, Photonics, and Optical Engineering (ICPPOE 2024), 2025:665–671.

[16] Xiao L, Yang X, Lan X, et al. Towards Visual Grounding: A Survey [J]. arXiv preprint arXiv:2412.20206, 2024.

[17] Gümüş, B., Balaban, M. Ö., & ÜNLÜSAYIN, M. (2011). Machine vision applications to aquatic foods: a review. Turkish Journal of Fisheries and Aquatic Sciences, 11(1).

[18] Antony, M. A., & Kumar, R. S. (2021, March). A comparative study on predicting food quality using machine learning techniques. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1771-1776). IEEE.

[19] Sun D-W. Inspecting pizza topping percentage and distribution by a computer vision method [J].Journal of food engineering, 2000, 44 (4): 245–249.

[20] Kuswantori A, Suesut T, Tangsrirat W, et al. Fish detection and classification for automatic sorting system with an optimized YOLO algorithm [J]. Applied Sciences, 2023, 13 (6): 3812.

[21] Li D, Wang Q, Li X, et al. Recent advances of machine vision technology in fish classification [J]. ICES Journal of Marine Science, 2022, 79 (2): 263–284.

[22] Yu X, Wang Y, Liu J, et al. Non-contact weight estimation system for fish based on instance seg mentation [J]. Expert systems with applications, 2022, 210: 118403.

[23] Zhang, L., Wang, J., & Duan, Q. (2020). Estimation for fish mass using image analysis and neural network. Computers and Electronics in Agriculture, 173, 105439.

[24] 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.

[25] Peng, B. (2019). Application of marine remote sensing technology in the development of fishery economy. Journal of Coastal Research, 94(SI), 783-787.

[26] Lim L-S L-S, Liew K-S, Yap T-K, et al. Length-weight relationship and relative condition factor of pearl oyster, Pinctada fucata martensii, cultured in the Tieshangang Bay of the Beibu Gulf, Guangxi Province, China [J]. Borneo Journal of Marine Science and Aquaculture (BJoMSA), 2020, 4 (1):24–27.

[27] Singh Y T. Relationships between environmental factors and biological parameters of Asian wedge clam, Donax scortum, morphometric analysis, length-weight relationship and condition index: a first report in Asia [J]. Journal of the Marine Biological Association of the United Kingdom, 2017,97 (8): 1617–1633.

[28] Yoshizumi, T., Gondolesi, G. E., Bodian, C. A., Jeon, H., Schwartz, M. E., Fishbein, T. M., ... & Emre, S. (2003, June). A simple new formula to assess liver weight. In Transplantation proceedings (Vol. 35, No. 4, pp. 1415-1420). Elsevier.

[29] Ji, X., Dahlgren, R. A., & Zhang, M. (2016). Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. Environmental monitoring and assessment, 188, 1-16.

[30] Liu, X., Du, K., Zhang, C., Luo, Y., Sha, Z., & Wang, C. (2023). Precision feeding system for largemouth bass (Micropterus salmoides) based on multi-factor comprehensive control. Biosystems Engineering, 227, 195-216.

[31] Harvey, B. C., & Railsback, S. F. (2007). Estimating multi-factor cumulative watershed effects on fish populations with an individual-based model. Fisheries, 32(6), 292-298.

[32] Dame R F. Comparison of various allometric relationships in intertidal and subtidal American oys ters [J]. Fishery Bulletin, 1972, 70 (4): 1121–1126.

[33] Pineda-Metz S E, Merk V, Pogoda B. A machine learning model and biometric transformations to facilitate European oyster monitoring [J]. Aquatic Conservation: Marine and Freshwater Ecosys tems, 2023, 33 (7): 708–720.

[34] Gimin R, Mohan R, Thinh L, et al. The relationship of shell dimensions and shell volume to live weight and soft tissue weight in the mangrove clam, Polymesoda erosa (Solander, 1786) from northern Australia [J]. NAGA, WorldFish Center Quarterly, 2004, 27 (3): 32–35.

[35] Vu S V, Knibb W, Nguyen N T, et al. First breeding program of the Portuguese oyster Crassostrea angulata demonstrated significant selection response in traits of economic importance [J]. Aquaculture, 2020, 518: 734664

[36] Vu S V, Knibb W, Gondro C, et al. Genomic prediction for whole weight, body shape, meat yield, and color traits in the Portuguese oyster Crassostrea angulata [J]. Frontiers in Genetics, 2021, 12:661276.

[37] Zhao, F., Hao, J., Zhang, H., Yu, X., Yan, Z., & Wu, F. (2024). Quality recognition method of oyster based on U-net and random forest. Journal of Food Composition and Analysis, 125, 105746.

[38] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.