Fish Species Classification Using an AquaFormerNet-Based Hybrid Deep Learning Approach

Chittireddy Ajith Reddy
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
ajithreddychittireddy@gmail.com

Chintha Pranay
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
chinthapranay127@gmail.com

Manne Akshitha
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
manneakshitha0201@gmail.com

M.A. Jabbar
Dept of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jabbar.meerja@gmail.com

***Abstract*—The identification of fish species is vital in fisheries management, food safety and nutritional awareness. Conventional forms of identification are time-consuming and need expert knowledge. This paper suggests a hybrid deep learning model, AquaFormerNet, to classify the species of the fish accurately using digital images. The suggested architecture will combine EfficientNet to extract features efficiently, a multi-head attention mechanism to concentrate on discriminative regions and a Transformer encoder to learn global relations among features.The dataset in this paper is a mixture of the Mendeley Fish Dataset (mxf2c45yb5) and other images obtained elsewhere, and includes about 7000+ images of 9 fish species. The pictures are superimposed and enhanced to enhance generalization. The model is trained with PyTorch framework with optimum hyperparameters.The proposed model has a better performance than the baseline models, and it has an accuracy of about 96%, as well as high values of precision, recall, and F1-score. Moreover, this proposed model offers information that is related to nutritional and health issues of the identified fish species, which makes it a whole-scale decision-support tool.**

Keywords: Fish Classification, Deep Learning, Computer Vision, EfficientNet, Vision Transformer, AquaFormerNet, Image Recognition, Artificial Intelligence

I. Introduction

The classification of fish species is a basic activity of fisheries management, marine biodiversity protection, environmental surveillance, and aquaculture. Proper detection of fish species can allow researchers and policymakers to examine the dynamics of populations, determine the well-being of the ecosystem, and achieve sustainable harvesting methods [5], [6]. Proper species identification in aquaculture is a prerequisite towards the optimal feeding, breeding and disease management which directly affect productivity and economic results. Nonetheless, the conventional approaches are highly dependent on manual classification by professionals such as marine biologists whereby morphological characteristics such as body shape, fin structure, coloration and scale patterns are studied [2]. These are effective, but are time consuming, specialised and subject to human error, particularly in cases of visual similar species, or large datasets.

As artificial intelligence and computer vision continue to develop at a high pace, automated image-based classification is stating its case as an attractive option. Deep learning models, especially Convolutional Neural Networks (CNNs) have been shown to perform well with the extraction of local features, including edges, textures, and patterns [1], [3]. Nonetheless, CNN-based methods usually do not discern global contextual relationships in an image, thus their use is obstructed in intricate situations. To solve this problem, attention-based and transformer architectures have been proposed, where models are able to pay attention to significant areas and learn long-range relationships [4], [9].CNNs + transformer mechanism Hybrid CNNs models have demonstrated superior performance as they are capable of learning both local and global features. In this study, we introduce AquaFormerNet, a hybrid deep learning framework based on EfficientNet, with attention and transformer modules to improve classification accuracy and efficiency [3], [10].

Although these improvements have been made, there are still some research gaps in the existing fish classification systems:

(i) Minimal adoption of hybrid architectures using CNNs and attention and transformer mechanisms.
(ii) Weakness in real world conditions like variations in light and background noise.
(iii) Lack of deployment-ready real-time systems to interact with users.

The current models tend to emphasize on the enhancement of feature extraction or that of classification accuracy, yet

they do not combine both local and global features learning effectively [3], [9]. Moreover, most methods are trained on controlled datasets and are not readily applicable to the real-world where images can differ in quality, lighting, and the complexity of the background [2], [5]. In addition, majority of research studies are restricted to experimental only kinds of research without offering easy to use deployment platforms thus limiting their applicability to real world applications [4], [6]. These challenges are necessary to establish an efficient, accurate and accessible system of fish species classification.

A. Objectives

The main Objectives of this paper are as follows:

1) Create an automated system of fish species classification based on deep learning methods of identifying images.
2) Develop and set up a hybrid model named AquaFormer-Net to extract and classify fish images and have precise features.
3) Process and analyze the fish image data set in order to enhance the quality and consistency of training data.
4) Implement the trained model in a web based system that enables users to post images of fish and receive a species prediction.
5) Determine the effectiveness of the proposed system in the criteria of accuracy, ease of use, and real-time predictions

B. Organization of Paper

This paper will be divided into a five sections. Section II shows the corresponding work, such as the traditional machine learning methods, deep learning-based models of fish classification, and the latest developments with attention and transformer-based models, as well as the comparison of existing models. Section III outlines the suggested system design and methodology with describing datasets, preprocessing methods, AquaFormerNet architecture, training settings, and the details of how the web-based application should be implemented. Section IV talks about the experimental findings, model performance metrics, confusion matrix analysis and model performance comparison with baseline models. Section V brings the paper to a close by summarizing the findings, pointing out limitations, and explaining possible future improvements.

II. Related Work

Initial studies in classifying fish species involved mainly the traditional machine learning methods with hand-crafted features, like color histograms, texture features, and shape features. These characteristics were added to classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Random Forests to carry out classification activities. Though these approaches were moderate in accuracy, their overall performance was greatly reliant on the selection of features and ineffective in complicated environments [5], [13]. Also, these methods were not able to make generalizations among different datasets because of different lighting, fish position, and background conditions.

As deep learning developed and became common, Convolutional Neural Networks (CNNs) found their way into fish classification because of their capability to automatically learn hierarchical feature representations. CNN-based models have proven to be very helpful in enhancing classification performance since they are able to extract local features like edges, textures, and patterns of fish images [11], [12]. Deep CNN architectures have been investigated in several studies to classify fish, such as multi-stage CNN pipelines, and transfer learning methods, which have shown better results than conventional methods. CNNs however primarily concentrate on the local feature extraction and tend to be poor at global contextual relationships in images, thus their use is constrained in differentiating between visually similar fish species.

To overcome these shortcomings, the recent studies have addressed the mechanisms of attention and transformer architectures to image classification tasks. Transformer models use self-attention mechanisms to learn long-range dependencies and global relations between features, which results in higher classification results [7], [8]. This has been demonstrated to be effective by hybrid models that incorporate CNNs and transformer architecture which combine both local and global feature learning [9], [10].Moreover, sophisticated CNN and hybrid deep learning networks have been created to enhance feature representation via discriminative parts of the fish image such as body structure, fin patterns, and texture variations. [14], [15]. These methods show how effective hybrid deep learning models are in enhancing classification of fish species.

Table I provides an overview of the recent research (2022-2025) on fish classification. One can note that deep CNNs models prevail over the older models, whereas hybrid CNN-transformer models are becoming the state of the art. Nevertheless, no overall systems integrating usability, robustness and accuracy in real-time are available.

TABLE I
Comparison of Fish Classification Studies

Ref No.	Model Used	Dataset	Accuracy
[5]	ML (SVM, RF)	Indian Fish Dataset	90.2%
[11]	CNN	Fish dataset	92.8%
[12]	Deep CNN (2-stage)	Fish dataset	94.2%
[14]	Attention-based CNN	Underwater fish dataset	95.0%
[9]	CNN + Transformer	Fish dataset	95.6%
[10]	Hybrid Transformer model	Marine dataset	96.1%
[7]	Transformer (ViT-based)	Image dataset	95.3%
[8]	Lightweight Transformer	Image dataset	94.8%
[15]	Attention-enhanced CNN	Fish dataset	95.2%
[13]	Traditional ML	Fish dataset	89.5%

III. Methodology

The proposed AquaFormerNet-based fish classification pipeline is shown in Fig. 1. A fish image will be inputted into the system, preprocessing will occur, and features will be

extracted using EfficientNet, followed by augmenting the feature representation using attention and transformer modules, and the ultimate classification output will be provided. The prediction is also displayed in the form of easily accessible web interface and other information about fish.

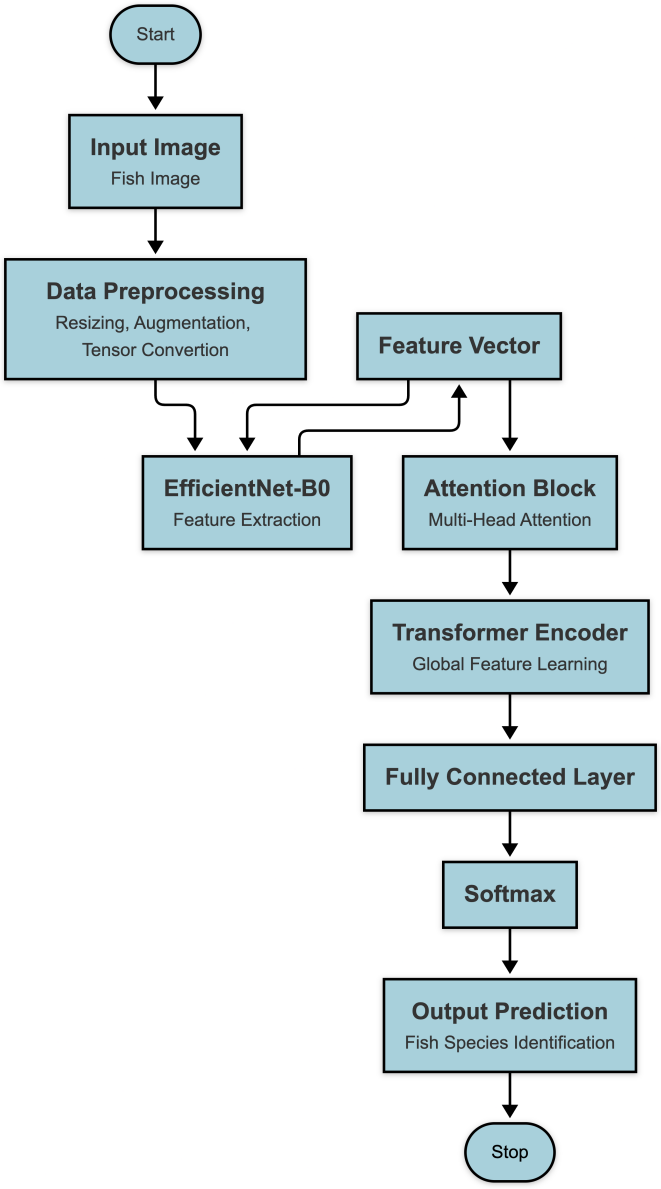

Fig. 1. Hybrid Fish Classification Architecture based on AquaFormerNet.

A. Datasets [5]

The model is trained on a mixture of publicly available data and custom-collected data. The main data is derived using the Mendeley Fish Dataset (mxf2c45yb5), which is supplemented with other fish images found in secondary sources to enhance the diversity and generalization.

Both the models are trained on a multi-classification task with 9 fish species. The data is divided into a structured format and is divided into training, validation and testing set in a stratified 70: 15:15 ratio. The overall dataset will have about 7000+ images, with enough level of representation of all classes shown in Table II. The dataset consists of variations in lighting conditions, orientations and backgrounds to approximate real-world conditions. To enhance the diversity of data and decrease the overfitting, data augmentation methods, including rotation, flipping, and scaling, will be implemented.

TABLE II
DATASET SUMMARY

Dataset	Classes	Samples	Split Ratio
Mendeley Fish Dataset + External Images [5]	9	7000+	70:15:15

B. Data Preprocessing

The input images are first subjected to some preprocessing stages before being pumped into the model. The steps are necessary to make the dataset consistent, offer a better training process, and increase the overall model performance in real-world conditions.

(i) *Resizing:* To adjust the input to the input demands of the EfficientNet-B0 architecture all input images are resized to a constant image size of 224 x 224 pixels. The dataset has images of different resolutions and aspect ratios, so resizing makes it similar and enables batch processing in training. The step also ensures that the computational complexity is kept down, without compromising on any significant visual attributes of the fish.

(ii) *Normalization:* The images are also normalized to a standard range so as to stabilize and speed up the training process. In most cases, the pixel values are normalized to a range of 0 to 1 or by the means and standard deviation of the values. This assists in diminishing the impact of the variability of illumination and makes the model converge more quickly in the optimization process.

(iii) *Data Augmentation:* In order to enhance the generalization power of the model and avoid overfitting, a number of data augmentation methods are used on the training data. These include:
Random Rotation: to deal with various orientation of fish. Horizontal Flipping: to model changes in mirrors. Scaling and Zooming: to take into consideration the size differences. Random Cropping: to enhance resistance to partial views.
These artificial enhancements in diversity of datasets and allow the model to work effectively in different real-world scenarios.

(iv) *Tensor Conversion:* Any preprocessed images are converted to a form as a tensor, which can be processed using deep learning systems like PyTorch. This procedure consists of converting the data of images to multi-dimensional arrays and organizing them in the order (Channels × Height × Width). Training and inference require conversion of tensors to efficiently be computed on GPUs.

(v) *Class Label Encoding:* Fish species have been labeled with distinct numbers to support supervised learning. The labels are then mapped to the indices of the classes, and the model is able to learn how input images relate to their categories.

(vi) *Data Splitting:* The dataset is stratified into training, validation and testing datasets to maintain a balanced distribution of classes. This helps in evaluating the model's performance on unseen data and prevents data leakage during training.

(vii) *Noise Handling and Quality Check:* Images that contain too much noise, are too blurred or contain some irrelevant background information are filtered or fixed to enhance the quality of the data. This is necessary in order to make sure that the model learns meaningful features and not noise.

C. Model Architecture and Training

The Hybrid deep learning architecture of AquaFormerNet is a hybrid of convolutional neural networks with attention and transformer mechanisms.

The architecture takes the form of the following components:

(i) *Feature Extraction (EfficientNet-B0):* EfficientNet is employed as the under-network in order to make rich features representations of input images. It generates a feature vector of size 1280, capturing important local features such as texture, shape, and patterns.

(ii) *Attention Block:* The extracted features are subjected to a multi-head attention mechanism which helps highlight the most important parts of the image shown in Table III e.g. fins, body structure, and texture patterns.

(iii) *Transformer Encoder:* The transformer encoder records worldwide dependencies and relationships among features and enhances the discerning capabilities of the model between visually similar fish species.

(iv) *Fully Connected Layer:* The processed features are then fed through a dense layer to do classification.

(v) *Softmax Layer:* A softmax function is used to give probability scores of each fish class to determine the final output.

TABLE III
MODEL CONFIGURATION

Component	Description
Backbone	EfficientNet-B0
Feature Dimension	1280
Attention	Multi-head Attention (8 heads)
Transformer Layers	2
Output Layer	Fully Connected + Softmax
Classes	9

The cross-entropy loss function is used to train the model and the Adam optimizer is used to optimize the model. Multiple epochs of training are done with an optimized batch size to support the use of a GPU. The overall classification accuracy of the proposed model is about 96%.

D. Training Strategy and Optimization

A number of training strategies are used in order to enhance the performance and stability of the model:

(i) *Transfer Learning:* Pretrained EfficientNet weights are utilized to realize faster convergence and enhance feature extraction.

(ii) *Regularization Techniques:* Overfitting is prevented by dropout and data augmentation.

(iii) *Learning Rate Optimization:* An appropriate learning rate is chosen to maintain stable learning and accelerated convergence.

(iv) *Evaluation Metrics:* Accuracy, precision, recall and F1-score are used to assess the model. Class-wise performance is evaluated with the help of a confusion matrix.

These measures will make sure that the model is accurate and at the same time it will be able to generalize to unseen data.

E. Web-Based Application Interface

The proposed system is deployed as an application which is web based so that it is both accessible and easy to use. Its application will enable the user to post images of fish and make immediate predictions.

The system is comprised of: An interface, created with HTML, CSS, and JavaScript and a modern glassmorphism design. A prediction server based on Flask to process prediction requests. An inference trained AquaFormerNet model.

The interface provides:

(i) Real-time fish species prediction

(ii) Score of the prediction.

(iii) Nutritional information and health benefits

The system is also set up to provide quick and precise predictions, and thus it can be used by students, researchers, fisheries professionals and general users.

IV. Results and Discussion

A. Overall Performance

The results of the proposed AquaFormerNet model on the held-out test dataset are provided in Table IV The model has a total classification accuracy of **96 %** which indicates that the model is effective in classifying several fish species in different conditions. This performance can be attributed to the hybrid architecture, a combination of EfficientNet to extract features and transformer-based attention to learn global context. The model also does not show any variation in the precision, recall, and F1-score values between the different classes (around 95%–96%), which shows that it does not favor certain classifications. The combination of attention mechanisms allows the model to center on discriminative areas like fins, body shape and texture enhancing the classification accuracy of visually similar fish species.

B. Evaluation Metrics

The performance of the proposed AquaFormerNet model is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score, which are widely used in image classification studies [12], [16]. These metrics are computed using the confusion matrix values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Accuracy represents the overall proportion of correctly classified samples. Precision indicates the percentage of positive predictions that are correct, while recall measures the ability of the model to identify all relevant samples. The F1-score provides a balanced measure by combining both precision and recall [17].

TABLE IV
HELD-OUT TEST SET PERFORMANCE OF AQUAFORMERNET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AquaFormerNet	96	96	95	95

C. confusion matrix and error analysis

The analysis of the confusion matrix Fig. 2 shows that most of the fish species are identified correctly, which is expressed through high values of the diagonal. There are however minor misclassifications between visually similar species like pomfret and black pomfret, tuna and mackerel. Such mistakes are created because of resemblances in tactile, form, and color schemes. In spite of these, the attention and transformer modules can alleviate these confusion by capturing the local and global feature dependencies.

D. Feature Importance and Heuristic Evaluation

The extraction of features in the proposed AquaFormerNet model is mainly processed by the EfficientNet-B0 backbone that captures fine-grained local features (edges, textures, and shape patterns of fish). These characteristics are further enhanced with the attention block that gives more significance to discriminative areas such as the fins, scales as well as body structure.

Transformer encoder is very important in learning international dependencies amid feature representations. This allows the model to identify contextual relationships in the image, and thus differentiate between visually similar fish species. It

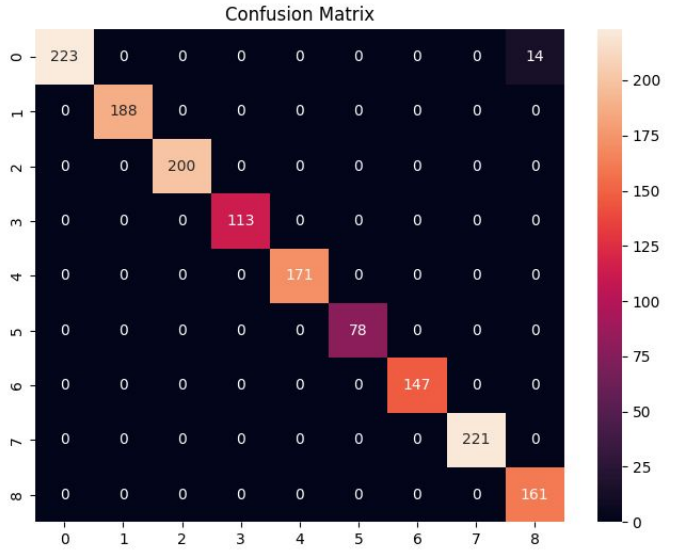

Fig. 2. Confusion table with true (actual) vs predicted class labels of the multi-class classification model.

has been experimentally observed that convolutional features combined with attention-based global learning have a big impact on enhancing the performance of classification. Of all components, EfficientNet backbone is associated with good feature representation, whereas the attention and transformer layers help to select features and understand the context. Such a hybrid solution provides high-performance in various lighting conditions, orientations and backgrounds

Our proposed system is implemented with an interactive web user interface, offering prediction and information outputs. Instant classification outputs, such as predicted fish class, and confidence level, are shown in Fig. 3. Other information such as nutritional values, health benefits and other information are illustrated in Fig. 4. These results show the value of the system in applications beyond typical image classification scenarios

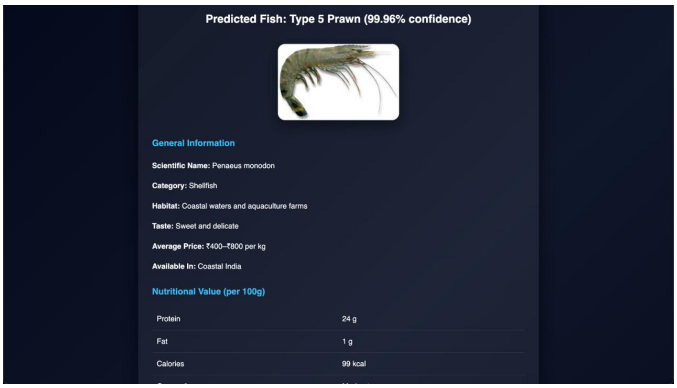

Fig. 3. Prediction interface of fish species with the uploaded image, the predicted class and the score of the confidence with the proposed model AquaFormerNet

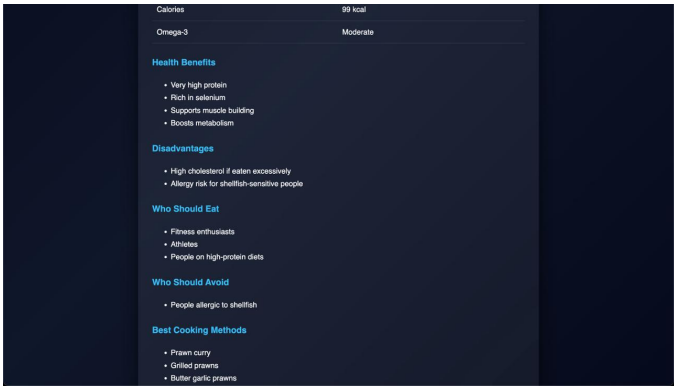

Fig. 4. Detailed information panel with nutritional, health benefits, and disadvantages, recommended consumption group, and cooking methods of the predicted fish species

E. Comparison with Baselines

The suggested AquaFormerNet model is contrasted with the current deep learning methods of fish classification. The more common CNN models, including ResNet, VGG, and models based on EfficientNets, are capable of reaching an accuracy of 8592 percent and 9495 percent respectively.

A hybrid AquaFormerNet architecture reaches a precision of **96%** and is better than these baseline models by combining both local feature extraction and global attention mechanisms. The proposed method, unlike the traditional models, offers enhanced discrimination between similar fish species and enhanced resilience against environmental changes.

V. Conclusion

In this paper, a hybrid deep learning framework that classifies fish species was introduced, AquaFormerNet, which integrates feature extraction via EfficientNet, attention, and transformer. The proposed method successfully reflects both the local and global characteristics and can discriminate various fish species. The model had an overall classification accuracy of about 96 % indicating excellent performance on different classes and image conditions. Combining the data augmentation and preprocessing methods also enhanced the robustness and generalization ability of the system.

Besides the classification model, a convenient web-based application had been created shown. which offered real-time predictions and specific nutritional and health-related information. Despite the promising results, the system continues to misclassify visually similar species, albeit with minor errors, and is sensitive to the quality of the dataset. The further development of work may be aimed at increasing the volume of data, enhancing the efficiency of the model, and extending the system with other possibilities, including the detection of fish freshness and the deployment of the system in the form of a mobile application. Altogether, the suggested system can serve as the effective and practical solution to the fish species classification in the real world.

References

[1] M. Aubard et al., "ROSAR: An Adversarial Re-Training Framework for Robust Side-Scan Sonar Object Detection," arXiv preprint, 2024.
[2] H. Zhou et al., "Real-time underwater object detection technology for complex underwater environments based on deep learning," Ecological Informatics, vol. 82, 2024.
[3] J. Ding et al., "Lightweight enhanced YOLOv8n underwater object detection network for low light environments," Scientific Reports, vol. 14, 2024.
[4] M. Vijayalakshmi and A. Sasithradevi, "AquaYOLO: Advanced YOLO-based fish detection for optimized aquaculture pond monitoring," Scientific Reports, 2025.
[5] P. Paygude et al., "Species identification for Indian seafood markets: A machine learning approach with a fish dataset," Data in Brief, vol. 58, 2025.
[6] Y. Cheng et al., "Efficient tuna detection and counting with improved YOLOv8 and ByteTrack in pelagic fisheries," Ecological Informatics, vol. 87, 2025.
[7] N. Carion et al., "End-to-End Object Detection with Transformers (DETR)," IEEE TPAMI, 2023 (extended works).
[8] Y. Lv et al., "RT-DETR: Real-time detection transformer," IEEE CVPR, 2023.
[9] X. Liu et al., "DP-FishNet: Dual-path transformer network for underwater detection," IEEE Transactions on Image Processing, 2024.
[10] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020, August. End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.
[11] Allken, V., Handegard, N.O., Rosen, S., Schreyeck, T., Mahiout, T. and Malde, K., 2019. Fish species identification using a convolutional neural network trained on synthetic data. ICES Journal of Marine Science, 76(1), pp.342-349.
[12] Jareño, J., Bárcena-González, G., Castro-Gutiérrez, J., Cabrera-Castro, R. and Galindo, P.L., 2024. Automatic labeling of fish species using deep learning across different classification strategies. Frontiers in Computer Science, 6, p.1326452.
[13] A. Shaikh et al., "An improved deep CNN-based freshwater fish classification with cascaded bio-inspired networks," *Automatika*, vol. 66, no. 2, pp. 249–280, 2025.
[14] A. Mohammadisabet, M. Karimi, and H. Rezaei, "CNN-Based Optimization for Fish Species Classification," *Information*, vol. 16, no. 2, p. 154, 2025.
[15] Hamzaoui, M., Rejili, M., Aoueileyine, M.O.E. and Bouallegue, R., 2025. DeepFishNET+: A Dual-Stream Deep Learning Framework for Robust Underwater Fish Detection and Classification. Applied Sciences, 15(20), p.10870.
[16] M. Mots'oehli, A. Nikolaev, W. B. IGede, J. Lynham, P. J. Mous, and P. Sadowski, "FishNet: Deep neural networks for low-cost fish stock estimation," arXiv:2403.10916, 2024.
[17] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers," IEEE/Pattern Recognition Metrics Study, 2022.