



An Explainable Multimodal Hybrid Deep Learning Framework for Real-Time IoT-Based Fish Freshness Classification

Rian Ordila¹, Miratu Megasari², Yulanda³, Yuda Irawan⁴, Refni Wahyuni⁵

¹Department of Information System, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia

²Department of Midwifery, Faculty of Health, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia

^{3,4,5}Department of Computer Science, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia

*rianordila@htp.ac.id

Abstract

Reliable assessment of fish freshness supports product quality, food safety, and consumer acceptance throughout fisheries supply chains. Sensory inspection remains common, but reliance on individual judgment can produce inconsistent assessments and limit monitoring efficiency. Here, a multimodal framework combines Computer Vision, IoT sensing, and Deep Learning to monitor Nile tilapia freshness in realtime. Its visual branch uses YOLO26 to locate fish eyes and EfficientNet B3 with CBAM to classify the extracted regions. A BiLSTM network learns temporal patterns from MQ136 and MQ137 measurements of H₂S and NH₃ associated with spoilage. An MLP then combines the class probabilities from both branches to assign the final freshness category. In the detection experiment, YOLO26 recorded 0.930 precision and 0.953 mAP@0.5. Visual classification with EfficientNet B3 plus CBAM reached 92% accuracy, whereas sensor classification with BiLSTM reached 93%, with AUC values as high as 0.99. Combining the modalities yielded 97% accuracy and a macro average F1 score of 0.97, exceeding the performance of either branch alone. Grad CAM maps located the main visual responses within eye regions relevant to freshness deterioration. These results support the framework's interpretability and its potential use for intelligent fish freshness assessment under near realtime monitoring conditions.

Keywords: BiLSTM; EfficientNet B3; fish freshness classification; multimodal fusion; YOLO26

How to Cite: R. Ordila, Miratu Megasari, Y. Yulanda, Yuda Irawan, and Refni Wahyuni, "An Explainable Multimodal Hybrid Deep Learning Framework for Real-Time IoT-Based Fish Freshness Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1349–1360, Oct. 2026, doi: [10.29207/resti.v10i5.7844](https://doi.org/10.29207/resti.v10i5.7844).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7844>

Received: May 13, 2026

Accepted: September 19, 2026

Available Online: October 7, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)

1. Introduction

Freshness determines several aspects of fish quality, including nutritional value, safety for consumers, selling price, and acceptability within distribution chains[1], [2]. In many developing countries, Nile tilapia (*Oreochromis niloticus*) is a common dietary source of protein because it is affordable and widely consumed[3], [4]. After harvest, tilapia can deteriorate quickly: its texture and color change, microorganisms proliferate, and hydrogen sulfide (H₂S) and ammonia (NH₃) are released as microbial activity and protein breakdown progress[5]. Assessors traditionally judge freshness by examining the eyes and gills, smelling the fish, and checking body texture. Such sensory judgments vary among observers and are difficult to apply consistently in large scale realtime monitoring[6], [7].

Developments in Artificial Intelligence (AI), Computer Vision, and the Internet of Things (IoT) have enabled automated freshness assessment with deep learning[8–10]. Research using Convolutional Neural Networks (CNN) shows that appearance and texture cues in fish photographs can distinguish freshness categories[11], [12]. EfficientNet offers effective feature extraction with relatively modest computational requirements. Vision Transformer (ViT) models and attention mechanisms have also strengthened visual representations used for classification[13], [14].

IoT sensing complements visual assessment by recording gases released as fish spoil[15], [16]. Measurements from MQ136 and MQ137 sensors provide information about H₂S and NH₃ associated with declining freshness[17]. Long Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks can analyze how these readings evolve over time, and

have shown promise for quality monitoring using sequential food sensor data[18], [19].

Current approaches leave several methodological gaps. Systems that rely on images or sensors alone can become less reliable when acquisition conditions or sensor readings change. Many visual studies classify whole fish photographs, although the eyes provide a particularly informative biological indicator of deterioration. Existing multimodal approaches often merge raw feature representations, increasing computation and retaining redundant information. Few tilapia monitoring frameworks bring together optimized object localization, visual attention, temporal gas sensing, fusion of class probabilities, and visual explanations within the same realtime IoT system. These gaps motivate a framework that uses localized eye information alongside the temporal evidence supplied by gas sensors.

The proposed framework connects YOLO26 eye localization with an EfficientNet B3 classifier enhanced by the Convolutional Block Attention Module (CBAM). A separate BiLSTM branch processes temporal measurements, and a Multilayer Perceptron (MLP) combines the resulting class probabilities. Cropping the eyes with the optimized YOLO26 detector directs visual learning toward regions that reveal freshness changes. In parallel, MQ136 and MQ137 sensors record realtime H_2S and NH_3 signals so that BiLSTM can learn the progression of spoilage. The fusion input contains six probability values, comprising three visual and three sensor outputs. This compact representation allows the MLP to learn how the two sources of evidence contribute to the final freshness decision.

The study contributes five connected elements: (1) optimized YOLO26 localization of tilapia eyes, (2) EfficientNet B3 classification with CBAM attention, (3) BiLSTM learning from temporal IoT measurements, (4) a lightweight MLP that fuses class probabilities, and (5) Grad CAM explanations of visual predictions. Together, these components connect biologically relevant image regions with the evolution of spoilage gas signals. The combined classifier reached 97% accuracy in the experiments, surpassing the unimodal baselines and supporting further development of intelligent realtime freshness monitoring.

2. Methods

Figure 1 outlines the development of the multimodal framework. The system links computer vision with IoT gas measurements and combines their class probabilities to produce an end to end deep learning assessment of fish freshness.

Figure 1 traces the two processing branches and their integration. YOLO26 first localizes the eyes, after which EfficientNet B3 with CBAM assigns visual probabilities (Pv) to Very Fresh, Fresh, and Not Fresh. The sensor branch receives MQ136 and MQ137

readings through an ESP32 module and uses BiLSTM to estimate sensor probabilities (Ps). Concatenating Pv and Ps creates a six dimensional input for the MLP fusion network, which returns the final class probabilities (Pf). Grad CAM provides heatmaps of the image regions that influence visual classification, allowing the system's freshness decisions to be examined during realtime monitoring.

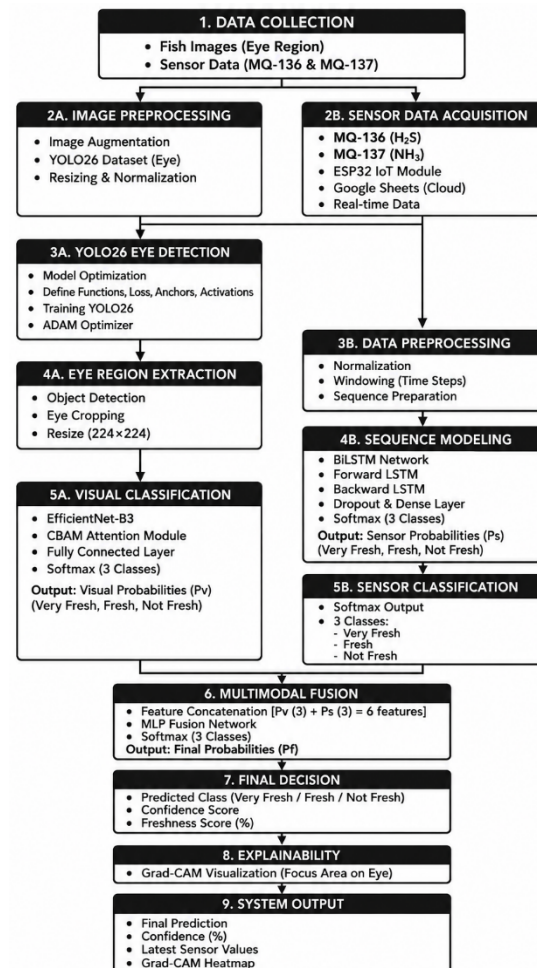


Figure 1. Flow of Model Development

2.1 Dataset Collection and Preprocessing

Visual data were obtained from Roboflow for Nile tilapia under different freshness conditions. Preparation produced two collections: 997 whole fish photographs for eye detection and 805 eye images for freshness classification. These collections represent appearance changes during deterioration, with particular attention to the eye as an informative quality indicator. Figure 2 displays examples used to develop the visual branch of the multimodal framework.

The examples in Figure 2 cover differences in illumination, viewing angle, and freshness. To standardize the inputs, Auto Orient corrected image orientation and each image was resized to 224×224 pixels for the EfficientNet B3 implementation. Horizontal flipping added variation during training and helped limit overfitting. Together, orientation correction, resizing, and augmentation improved input

consistency while exposing the classifier to a wider range of visual appearances.

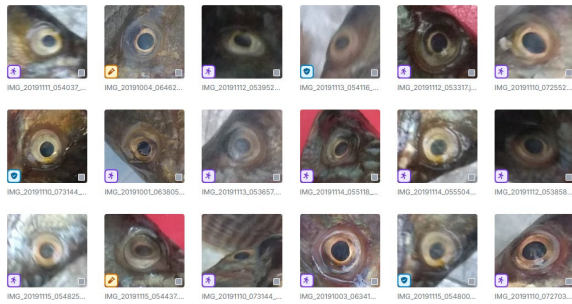


Figure 2. Sample Visualization of Nile Tilapia Eye Image Dataset

Data preparation began with 997 whole fish photographs. Annotation yielded 1,086 eye instances because individual photographs could contain multiple visible eyes. Detection and cropping were followed by curation, producing 805 eye images for visual classification. The IoT system separately supplied 1,000 sensor records for temporal analysis. Each dataset was partitioned into training, validation, and test subsets. Results tables report only the test samples relevant to the corresponding task, so their sample counts differ from the acquisition totals. This distinction reflects the separate detection, visual classification, and sensor analysis stages.

The IoT monitoring system supplied 1000 records from MQ136 and MQ137 sensors across different freshness conditions. These measurements supported temporal modeling and the sensor component of multimodal fusion. Table 1 gives examples of the resulting sensor dataset.

Table 1. Sample Sensor Dataset for Fish Freshness Classification

MQ136 (PPM)	MQ137 (PPM)	Label
0.27	1.69	Very Fresh
0.48	1.55	Very Fresh
0.68	2.56	Fresh
0.64	3.47	Fresh
1.30	5.26	Not Fresh
0.38	2.08	Very Fresh
0.29	1.65	Very Fresh
1.03	3.21	Fresh

Table 1 pairs measurements from MQ136 and MQ137 with the Very Fresh, Fresh, and Not Fresh labels. MQ136 responds to H₂S, whereas MQ137 measures ammonia related signals associated with spoilage. Before modeling, the readings were normalized to reduce differences in feature scale and organized into sequences for BiLSTM. This preparation improved input consistency and stabilized temporal learning. The numbers in Table 1 are normalized sensor measurements used as model features; they do not represent raw gas concentrations or absolute ppm values. Their interpretation therefore concerns relative, scaled sensor responses rather than direct chemical concentration measurements.

Ground truth assignment followed a controlled protocol combining post harvest storage duration with visual inspection. Assessors considered eye transparency,

shape, brightness, and the fish's general appearance. Clear, bright eyes indicated Very Fresh samples; moderate appearance changes indicated Fresh samples; and obvious deterioration, including cloudy eyes, indicated Not Fresh samples. The same protocol governed labeling for images and sensor records to maintain annotation consistency and reduce subjective variation.

Partitioning preceded model development to support reproducibility and limit data leakage. Stratified sampling divided the eye image and sensor datasets into training, validation, and test subsets while preserving the distribution of Very Fresh, Fresh, and Not Fresh. Validation samples supported parameter selection and training monitoring; independent test samples were reserved for the final evaluation. A random seed of 42 was fixed for dataset partitioning and model initialization across experiments. These procedures maintained separate roles for the data used to learn, select, and evaluate the models.

2.2 YOLO26 Eye Detection and Extraction

YOLO26, implemented through Ultralytics, located fish eyes before visual classification. Its suitability for realtime edge detection motivated this choice, including end to end inference without NMS, Progressive Loss Balancing (ProgLoss), Small Target Aware Label Assignment (STAL), and lightweight deployment optimizations[20], [21]. These properties support localization of small eye regions despite changes in lighting and acquisition conditions. Training used one object category, eye, and Adam optimization for 100 epochs with an initial learning rate of 0.001. Resizing, horizontal flipping, and normalization prepared the training inputs and supported generalization. Computation was performed in a GPU environment using the Ultralytics framework. Detected eye regions were cropped, resized to 224×224 pixels, and passed to EfficientNet B3 plus CBAM. Equation 1 expresses the detection objective.

$$\mathcal{L}_{YOLO26} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{obj}\mathcal{L}_{obj} \quad (1)$$

As shown in equation 1, $\mathcal{L}_{YOLO26}$ represents the total loss function of the YOLO26 model, $\mathcal{L}_{box}$ denotes the bounding box regression loss used for object localization, $\mathcal{L}_{cls}$ represents the classification loss for object category prediction, and $\mathcal{L}_{obj}$ corresponds to the objectness confidence loss indicating the probability of object existence within the predicted region. Meanwhile, λ_{box} , λ_{cls} , and λ_{obj} are weighting coefficients used to balance each optimization component during training. In addition, the final bounding box prediction generated by YOLO26 can be represented as Equation 2.

$$B = (x, y, w, h, c) \quad (2)$$

In Equation 2, the detected eye's center is specified by x and y, its bounding box dimensions by w and h, and the detection confidence by c. Localizing and cropping the eyes enables classification to concentrate on

appearance cues associated with freshness deterioration. This step also reduces the influence of surrounding image content and the computation required for subsequent visual analysis.

YOLO26 was chosen for its compact architecture and applicability to edge detection. Its main role was accurate eye localization, allowing the subsequent freshness classifier to process a focused region with less background interference and lower computational demand.

2.3 Visual Classification

Following YOLO26 localization, the cropped eyes entered a visual classifier combining EfficientNet B3 with CBAM. This combination was designed to refine

discriminative image features and improve classification reliability. Figure 3 depicts the architecture used to assign freshness categories.

Figure 3 organizes visual classification into an input layer, the EfficientNet B3 backbone, CBAM, and a fully connected output layer[22], [23]. Each $224 \times 224 \times 3$ image passes through a stem convolution and a sequence of MBConv blocks. MBConv1 and MBConv6 operations use 3×3 and 5×5 kernels to build increasingly abstract eye representations. Channel and spatial attention in CBAM then refine the backbone feature map, strengthening responses to relevant freshness cues and reducing the contribution of less informative features.

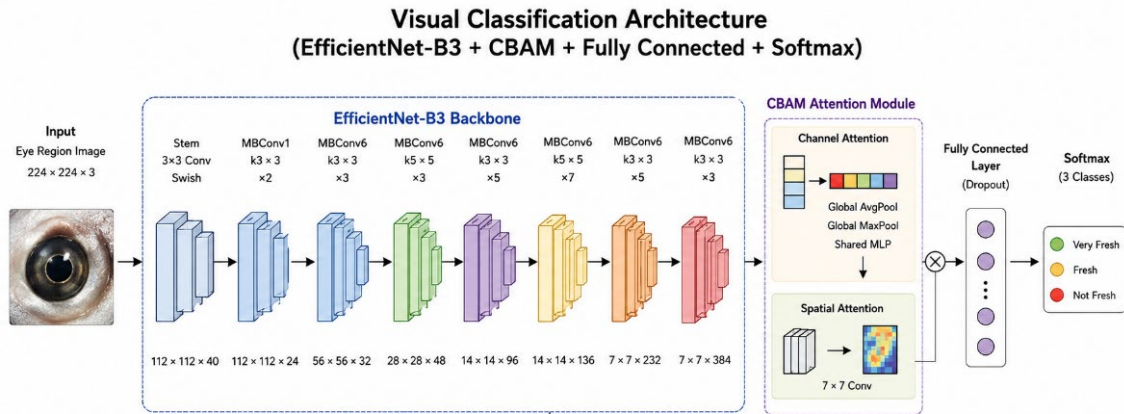


Figure 3. EfficientNet B3 and CBAM Based Visual Classification Architecture

Mathematically, the CBAM attention mechanism can be expressed as Equations 3-4.

$$F' = M_c(F) \otimes F \quad (3)$$

$$F'' = M_s(F') \otimes F' \quad (4)$$

As defined in Equations 3 and 4, F represents the input feature map extracted by EfficientNet B3, $M_c(F)$ denotes the channel attention map, $M_s(F')$ represents the spatial attention map, and $\otimes$ indicates element wise multiplication. The channel attention module focuses on important feature channels using global average pooling and global max pooling operations, while the spatial attention module identifies important spatial regions using a 7×7 convolution operation. After the attention refinement process, the extracted features are forwarded into the fully connected layer followed by the Softmax activation function to generate three visual probability outputs (Pv), corresponding to the classes Very Fresh, Fresh, and Not Fresh. The integration of EfficientNet B3 and CBAM constitutes one of the proposed model novelties because it enables efficient feature extraction with adaptive attention enhancement for fish freshness classification based specifically on eye region characteristics.

2.4 IoT Sensor Data Acquisition

The acquisition hardware connects gas sensing, camera input, wireless communication, and cloud storage for realtime multimodal monitoring. Figure 4 shows how

these devices and services exchange the data used for freshness assessment.

In Figure 4, MQ136 and MQ137 sensors operate alongside a fish eye camera and a Wemos ESP32 microcontroller. Google Sheets provides cloud synchronization of the acquired data. MQ136 records H_2S related signals, and MQ137 records ammonia related signals arising during spoilage. Sensor readings and image information are transmitted to web and mobile dashboards for realtime viewing and multimodal freshness analysis.

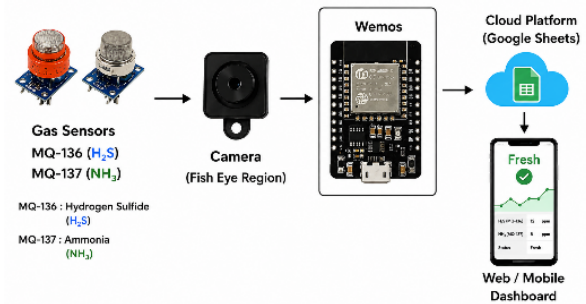


Figure 4. IoT Based Fish Freshness Data Acquisition System

Both gas sensors were allowed to stabilize before acquisition to limit fluctuations during startup. The ESP32 sampled readings continuously at fixed intervals and stored them for subsequent analysis. Normalization and noise reduction were applied to the raw values

before training. These preprocessing operations aimed to improve measurement consistency by reducing signal instability and the effects of drift and environmental variation commonly encountered with MQ-series sensors.

2.5 Sequence Modeling and Sensor Classification

The sensor branch uses BiLSTM to learn how MQ136 and MQ137 gas measurements change as freshness declines. Processing each sequence in both directions captures temporal relationships that support classification[24]. Table 2 lists the architecture selected for this sensor model.

Table 2 specifies a Bidirectional LSTM layer producing 128 features, a Dropout layer, and a Dense hidden layer containing 32 neurons. A final Dense layer returns three outputs for Very Fresh, Fresh, and Not Fresh.

Table 2. BiLSTM Architecture for Sensor Classification

Layer (Type)	Output Shape	Parameters
Bidirectional (BiLSTM)	(None, 128)	34,304
Dropout	(None, 128)	0
Dense	(None, 32)	4,128
Dense Output Layer	(None, 3)	99

BiLSTM reads sensor sequences in forward and backward order, allowing its representation to capture temporal relationships within spoilage related gas fluctuations. Equation 5 defines the resulting hidden state.

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \quad (5)$$

As defined in Equation 5, h_t represents the final hidden state at time step t , $\vec{h}_t$ denotes the forward hidden state, $\overleftarrow{h}_t$ indicates the backward hidden state, and $\oplus$ represents the concatenation operation between both directional outputs. Furthermore, the LSTM memory update process is formulated as Equation 6.

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

In Equation 6, c_t is the current cell state, f_t denotes the forget gate, i_t represents the input gate, $\tilde{c}_t$ is the candidate memory state, and $\odot$ indicates element wise multiplication. The integration of BiLSTM in the proposed framework constitutes an important component because it enables adaptive temporal sequence learning from realtime gas sensor signals before generating sensor probabilities (Ps) for the multimodal fusion stage.

2.6 Multimodal Fusion

An MLP joins the class probabilities returned by EfficientNet B3 plus CBAM and BiLSTM. By learning relationships between the visual and sensor predictions, it produces a combined probability distribution for the final freshness decision. Figure 5 illustrates this fusion architecture.

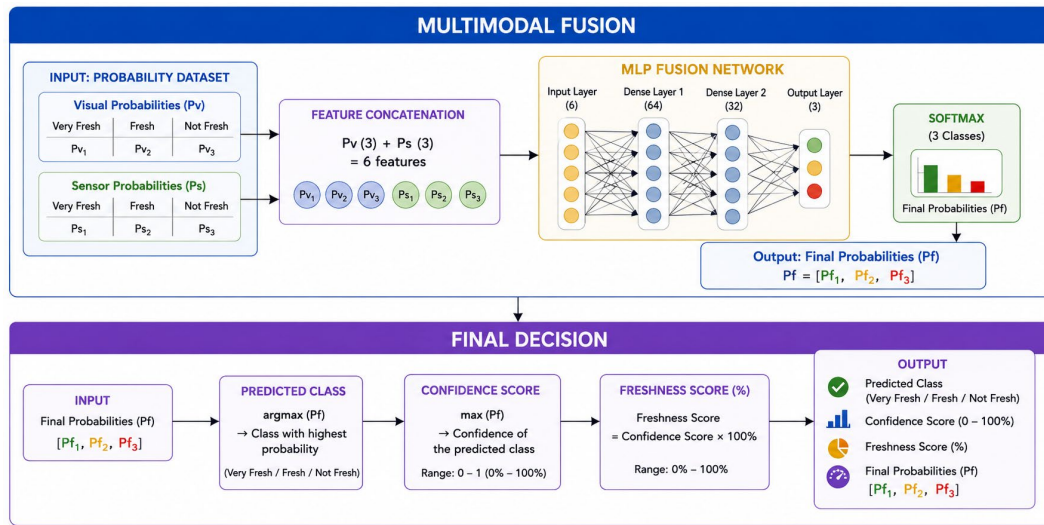


Figure 5. Proposed Multimodal Fusion and Final Decision Architecture

Referring to Figure 5, the proposed multimodal fusion framework combines visual probabilities $P_v = [P_{v1}, P_{v2}, P_{v3}]$ and sensor probabilities $P_s = [P_{s1}, P_{s2}, P_{s3}]$ into a six dimensional feature vector through feature concatenation. The concatenated probability vector is mathematically represented as Equation 7.

$$F = P_v \oplus P_s \quad (7)$$

F denotes the multimodal feature vector, P_v represents the visual probability outputs from the EfficientNet B3

plus CBAM model, P_s denotes the sensor probability outputs from the BiLSTM model, and $\oplus$ indicates the concatenation operation. The resulting six dimensional vector is then processed using an MLP fusion network consisting of multiple dense layers and dropout regularization layers to learn nonlinear relationships between visual and sensor modalities. The final prediction probabilities are generated using the Softmax activation function formulated as Equation 8.

$$Pf_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}} \quad (8)$$

Pf_i represents the final probability for class i , z_i denotes the output logit value of class i , and C indicates the total number of classes. The proposed probability level fusion strategy constitutes the main novelty of this study because previous multimodal fish freshness studies commonly perform fusion at raw data or feature extraction levels, whereas the proposed framework integrates probabilistic confidence information adaptively to improve classification robustness and decision consistency. The MLP fusion network architecture used in this study consists of multiple dense and dropout layers for multimodal probability learning. Table 3 explains the detailed architecture configuration of the proposed MLP fusion model.

As detailed in Table 3, the MLP contains Dense hidden layers with 32 and 16 neurons, with Dropout used to limit overfitting. The output Dense layer applies Softmax to generate probabilities for Very Fresh, Fresh, and Not Fresh. Only 803 parameters are trainable, keeping the probability fusion network compact and computationally economical for potential realtime deployment.

Table 3. MLP Architecture for Multimodal Fusion

Layer (Type)	Output Shape	Parameters
Dense Layer 1	(None, 32)	224
Dropout Layer 1	(None, 32)	0
Dense Layer 2	(None, 16)	528
Dropout Layer 2	(None, 16)	0
Dense Output Layer	(None, 3)	51

2.7 Explainability AI (XAI)

Gradient weighted Class Activation Mapping (Grad CAM) was used for the Explainability Artificial Intelligence (XAI) analysis. Applied to the final convolutional layer of EfficientNet B3 plus CBAM, it produces heatmaps of the eye regions influencing freshness predictions. Equation 9 defines the class activation calculation used to examine these visual responses.

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (9)$$

$L_{GradCAM}^c$ represents the class activation map for class c , A^k denotes the activation feature map, and α_k^c represents the importance weight computed from the gradient of the target class score. The integration of Grad CAM improves model transparency and interpretability by visually explaining the prediction process of the proposed fish freshness classification framework.

2.8 Finalization and Evaluation of the Model

Evaluation covered both localization and freshness classification. For YOLO26, the measures were mean Average Precision (mAP), precision, recall, and confidence score. EfficientNet B3 plus CBAM, BiLSTM, and MLP fusion were assessed using accuracy, precision, recall, F1 score, confusion matrices, ROC curves, and Area Under Curve (AUC) [25-27]. Independent test data supplied every reported performance result; validation data supported model

selection and monitoring only. This separation provided a consistent basis for assessing the detection and classification branches. Formal hypothesis tests were outside the study's scope and remain a task for future validation.

3. Results and Discussions

3.1 Performance of YOLO26 Model Training Results

Training optimized YOLO26 for eye localization before classification. Loss trajectories, precision, recall, and mean Average Precision (mAP) were examined to assess convergence and detection performance. Figure 6 plots these training measurements.

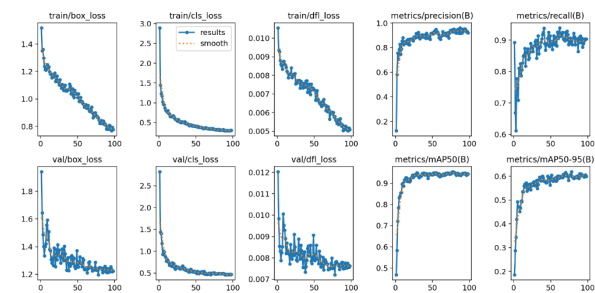


Figure 6. YOLO26 Training Performance Results

Across the 100 epochs in Figure 6, box, classification, and distribution focal losses decline steadily, showing convergence of the detector. Precision and recall during training exceed 0.90, while mAP50 and mAP50-95 approach 0.95 and 0.80, respectively. These trajectories suggest effective localization under the image conditions represented in training. The limited separation between training and validation losses also indicates that overfitting was restrained, supporting generalization to unseen eye images.

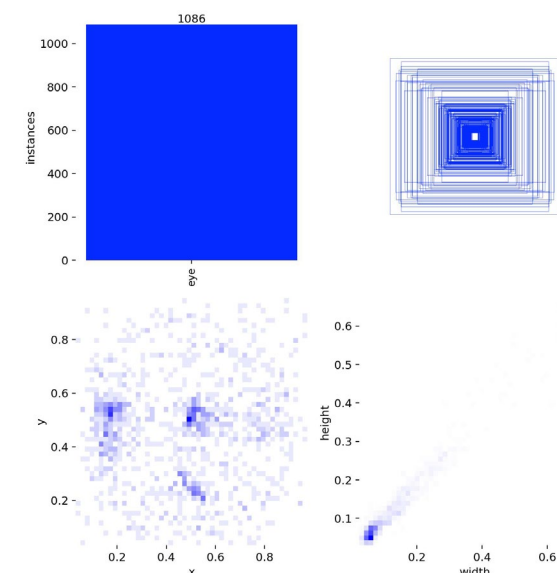


Figure 7. Fish Eye Annotation Distribution and Bounding Box Analysis

Before training YOLO26, the annotations were examined for object representation, location

consistency, and bounding box geometry. Figure 7 summarizes the label distribution and the spatial characteristics of the detection dataset.

Figure 7 summarizes 1,086 annotated eyes within a single detection category. The spatial heatmap places most eyes near image centers, reflecting fairly consistent positioning during acquisition. Box width and height distributions show that the eyes occupy small, proportionally bounded regions, making small object localization relevant to this task. Table 4 reports the quantitative detection evaluation used to assess the YOLO26 model's ability to identify and localize these eye regions.

Table 4. YOLO26 Model Performance Evaluation Results

Class	Images	Instances	Precision (Box(P))	Recall (R)	mAP50	mAP50-95
All	185	278	0.93	0.896	0.953	0.619

The test results in Table 4 give YOLO26 a precision of 0.93 and recall of 0.896, indicating reliable eye identification with relatively few missed objects. Its mAP50 is 0.953 and mAP50-95 is 0.619, summarizing localization performance across Intersection over Union (IoU) thresholds. These findings support the detector's use within freshness monitoring. Although its compact architecture offers potential for computationally efficient deployment, latency and throughput were not measured in this study. Suitability for near realtime operation therefore remains a practical prospect requiring dedicated measurement. Table 4 reports the held out detection subset: 185 images containing 278 annotated eyes. These samples were reserved after partitioning and therefore represent only the evaluation data, rather than the full acquired collection.

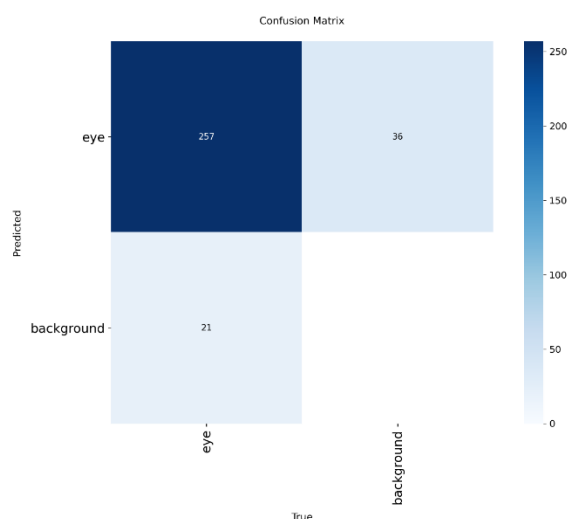


Figure 8. Confusion Matrix of YOLO26 Detection Model

The confusion matrix in Figure 8 records 257 correctly detected eyes, together with 36 false positives and 21 false negatives. Background complexity and variations in object appearance still produced detection errors. The 278 annotated eyes comprise the 257 true positives and

21 missed detections. The 36 additional predictions are false positives without corresponding annotated eyes. Figure 9 illustrates examples of the detector's bounding box outputs.



Figure 9. Sample Results of YOLO26 Fish Eye Detection

Figure 9 illustrates eye localization across fish species, orientations, and acquisition environments. Bounding boxes generally align with the visible eye regions, including small eyes against cluttered backgrounds or under partial occlusion. These examples support the detector's robustness within the proposed monitoring workflow.

3.2 Classification Results Using EfficientNet-B3

The baseline EfficientNet B3 classifier received eye crops from YOLO26 and assigned freshness categories. Figure 10 shows its class predictions and error distribution. Correct assignments comprised 31 Very Fresh, 20 Fresh, and 19 Not Fresh samples. Three Very Fresh samples were assigned to Fresh, while two samples from each of the Very Fresh and Fresh groups were assigned to Not Fresh. The distribution suggests that the backbone extracted useful eye features across the categories while retaining some classification ambiguity. In total, the matrix covers 81 test images, matching the supports in Table 5: 36 Very Fresh, 23 Fresh, and 22 Not Fresh.

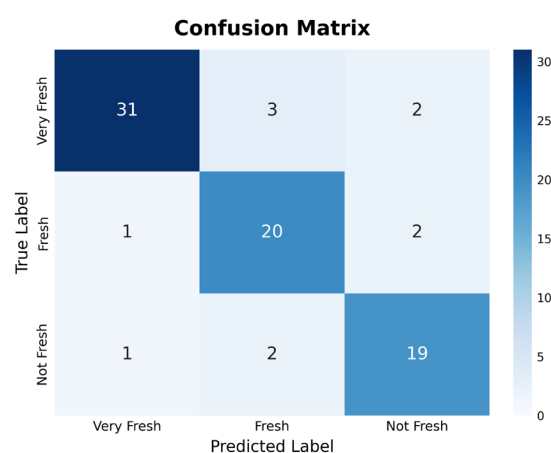


Figure 10. Confusion Matrix of EfficientNet B3 Classification Model

Table 5 summarizes EfficientNet B3 classification using precision, recall, F1 score, and accuracy. Evaluation used the visual test subset of 81 eye images,

distributed among the three freshness categories, to examine consistency of the class predictions.

Table 5. Classification Performance of EfficientNet B3 Model

Class	Precision	Recall	F1-Score	Support
Very Fresh	0.94	0.86	0.90	36
Fresh	0.80	0.87	0.83	23
Not Fresh	0.83	0.86	0.84	22
Accuracy		0.86		81
Macro Average	0.86	0.86	0.86	81
Weighted	0.87	0.86	0.87	81
Average				

For EfficientNet B3, Table 5 gives the Very Fresh category the highest precision, 0.94, reflecting few false positive assignments to that class. Overall accuracy is 0.86 and the weighted average F1 score is 0.87. Recall reaches 0.87 for Fresh and 0.86 for Not Fresh, indicating broadly comparable performance for these categories.

3.3 Classification Results Using EfficientNet-B3 + CBAM

CBAM was added to EfficientNet B3 to emphasize eye features informative of freshness. The resulting classifier was examined through its confusion matrix to assess class separation and prediction errors. Figure 11 displays the outcomes after attention refinement.

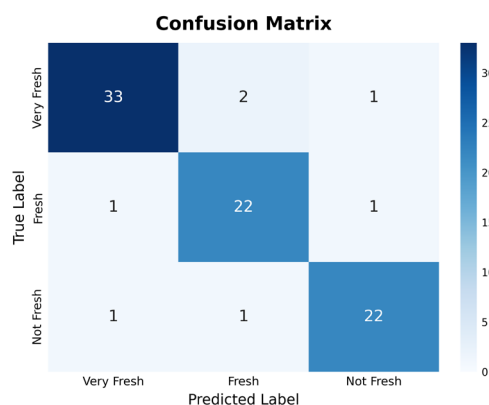


Figure 11. Confusion Matrix of EfficientNet B3 + CBAM Model

Figure 11 records 33 correct Very Fresh predictions and 22 correct predictions for each of Fresh and Not Fresh. Of the Very Fresh samples, two were assigned to Fresh and one to Not Fresh. Each of the other categories also had two classification errors. Relative to the baseline, the pattern suggests clearer separation of freshness categories after CBAM refinement. The 84 evaluated images comprise 36 Very Fresh, 24 Fresh, and 24 Not Fresh samples, matching the class supports in Table 6. That table summarizes classification after eye extraction and attention based visual processing.

Table 6 reports 0.92 accuracy for EfficientNet B3 plus CBAM, with weighted precision, recall, and F1 score also equal to 0.92. Very Fresh attains the highest class F1 score, 0.93. Recall for both Fresh and Not Fresh is 0.92, supporting consistent feature discrimination across the three categories following attention refinement.

Table 6. Classification Performance of EfficientNet B3+CBAM Model

Class	Precision	Recall	F1-Score	Support
Very Fresh	0.94	0.92	0.93	36
Fresh	0.88	0.92	0.90	24
Not Fresh	0.92	0.92	0.92	24
Accuracy		0.92		84
Macro Average	0.91	0.92	0.91	84
Weighted	0.92	0.92	0.92	84
Average				

ROC analysis examined the class separation provided by EfficientNet B3 plus CBAM over varying decision thresholds. Figure 12 presents the three class curves used for this assessment.

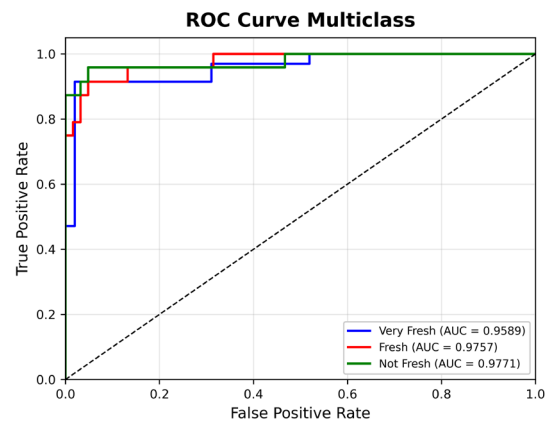


Figure 12. Multiclass ROC Curve of EfficientNet B3 + CBAM Model

In Figure 12, AUC is 0.9589 for Very Fresh, 0.9757 for Fresh, and 0.9771 for Not Fresh. The curves approach the graph's upper left region, reflecting high sensitivity with relatively few false positives across thresholds. AUC exceeds 0.95 for every category, supporting the attention enhanced classifier's ability to distinguish visual freshness cues.

Figure 13 compares evaluation metrics for baseline EfficientNet B3 and its CBAM variant. This comparison examines the empirical contribution of attention to visual freshness classification.

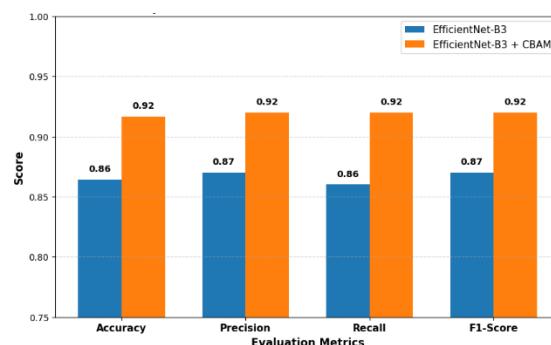


Figure 13. Performance Comparison Between EfficientNet B3 and EfficientNet B3 + CBAM

Figure 13 favors EfficientNet B3 plus CBAM across the reported metrics. Accuracy rises from 0.86 to 0.92, while precision and F1 score rise from 0.87 to 0.92. These changes suggest that attention refinement

improved the representation of eye features and reduced prediction ambiguity. Since the study did not test statistical significance, the differences describe observed performance gains rather than statistically established superiority. Grad CAM was applied to examine which visual regions influenced EfficientNet B3 plus CBAM predictions. Figure 14 displays the resulting heatmaps for the freshness classification examples.

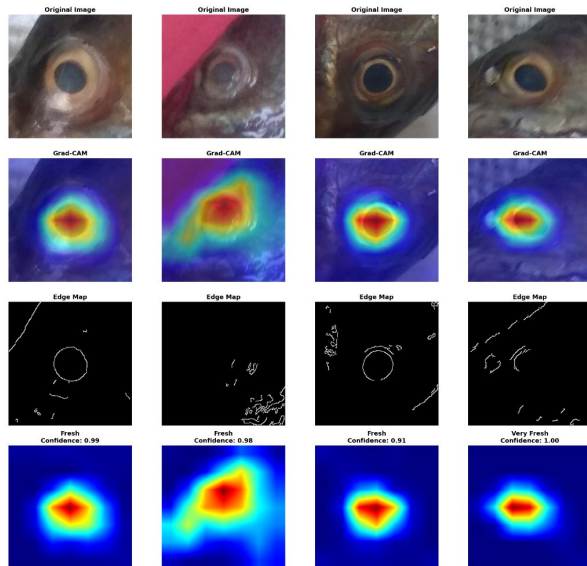


Figure 14. Grad CAM Visualization of Fish Eye Freshness Classification

The heatmaps in Figure 14 place the strongest responses around the eye center, pupil, and surrounding texture. Several examples have prediction confidence between 0.91 and 1.00. This concentration indicates that the classifier relies on eye features relevant to freshness, and supports the role of CBAM in directing visual responses away from background content.

3.4 Classification Results Using BiLSTM

BiLSTM classified the temporal measurements supplied by MQ136 and MQ137. The complete sensor dataset contained 350 Very Fresh, 350 Fresh, and 300 Not Fresh samples.

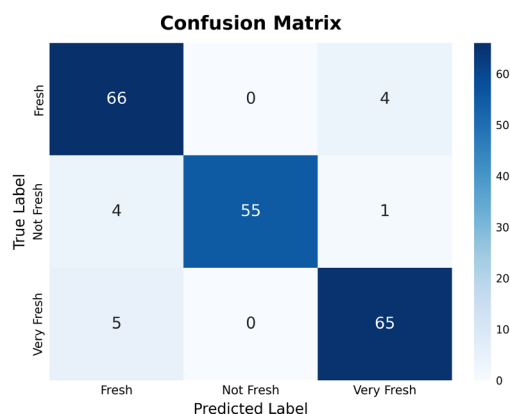


Figure 15. Confusion Matrix of BiLSTM Sensor Classification Model

Figure 15 shows the confusion matrix used to assess how consistently the model assigned classes from sensor sequences. Figure 15 records 66 correct Fresh predictions, 55 correct Not Fresh predictions, and 65 correct Very Fresh predictions. The relatively few errors support BiLSTM's ability to learn temporal freshness information from the IoT readings. Evaluation used an independent subset of 200 sensor sequences. Table 7 summarizes the associated classification metrics.

Table 7. Classification Performance of the BiLSTM Sensor Classification Model

Class / Metric	Precision	Recall	F1-Score	Support
Fresh	0.88	0.94	0.91	70
Not Fresh	1.00	0.92	0.96	60
Very Fresh	0.93	0.93	0.93	70
Accuracy		0.93		200
Macro Average	0.94	0.93	0.93	200
Weighted Average	0.93	0.93	0.93	200

As shown in Table 7, BiLSTM reaches 0.93 accuracy, with macro precision, recall, and F1 score above 0.93. Not Fresh has precision of 1.00 and F1 of 0.96; recall is 0.94 for Fresh and 0.93 for Very Fresh. Normalization and other preprocessing reduced variability in the MQ136 and MQ137 signals, helping the network learn temporal deterioration patterns despite potential environmental effects on sensor responses.

Figure 16 examines BiLSTM class discrimination through multiclass ROC analysis. AUC reaches 0.97 for Fresh, 0.98 for Not Fresh, and 0.99 for Very Fresh. Curves near the upper left of the plot reflect high true positive rates relative to false positives. These results support the network's ability to distinguish freshness categories from temporal gas sensor measurements.

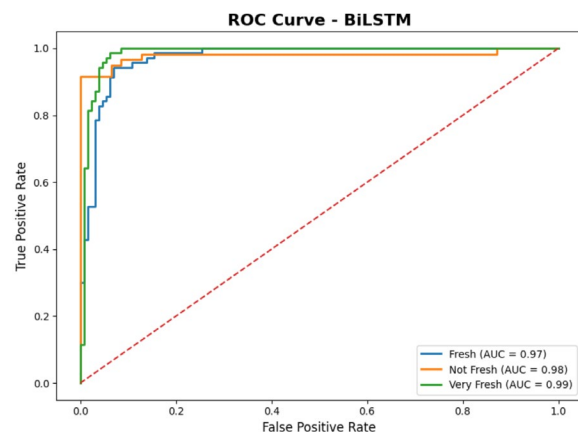


Figure 16. Multiclass ROC Curve of BiLSTM Model

3.5 Multimodal Visual Integration and IoT Sensor

Integration combines the visual class probabilities from EfficientNet B3 plus CBAM with those from the BiLSTM sensor branch. The resulting probability representation supplies the MLP, which learns the final freshness decision from both modalities.

Table 8 illustrates the combined class probabilities supplied by the visual and sensor branches.

Table 8. Sample Multimodal Probability Integration Results

vis veryfresh	vis fresh	vis notfresh	sen veryfresh	sen fresh	sen notfresh	label
0.9845	0.0141	0.0014	0.0068	0.9519	0.0413	0
0.4565	0.5430	0.0006	0.1494	0.8285	0.0222	1
0.0163	0.0075	0.9761	0.8307	0.1642	0.0051	2
0.0052	0.0012	0.9936	0.0250	0.9628	0.0122	2
0.9996	0.0003	0.0001	0.0017	0.9726	0.0257	0

Table 8 includes several class probabilities above 0.95, showing confident outputs from the individual branches. The examples also reveal complementary information: sensor evidence can support a decision when visual confidence is lower. This behavior motivates learning a joint decision from both probability distributions. Table 9 reports 0.97 accuracy for the fusion model, with macro precision, recall, and F1 score of 0.97. Precision for Not Fresh is 1.00 and its F1 score is 0.98; recall is 0.97 for Very Fresh and 0.96 for Fresh.

Table 9. Classification Performance of the Multimodal Fusion Model

Class	Precision	Recall	F1-Score	Support
Very Fresh	0.97	0.97	0.97	39
Fresh	0.92	0.96	0.94	25
Not Fresh	1.00	0.96	0.98	26
Accuracy		0.97		90
Macro Average	0.97	0.97	0.97	90
Weighted Average	0.97	0.97	0.97	90

Figure 17 displays the confusion matrix used to examine the fusion model's class assignments.

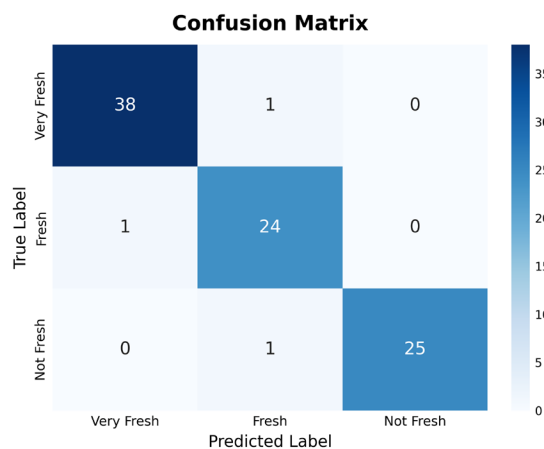


Figure 17. Confusion Matrix of the Multimodal Fusion Model

The fusion matrix in Figure 17 contains 38 correct Very Fresh assignments, 24 correct Fresh assignments, and 25 correct Not Fresh assignments. Only three samples are misclassified overall, yielding approximately 97% accuracy. These results suggest that combining image and sensor evidence reduces ambiguity and improves prediction consistency. Figure 18 further examines the multimodal MLP through its multiclass ROC curves.

Figure 18 gives AUC values of 0.98 for Very Fresh and Fresh, and 1.00 for Not Fresh. The curves lie near the upper left of the plot, indicating strong discrimination with relatively low false positive rates. The results support MLP fusion of the visual and sensor

probabilities. Figure 19 presents deployment of the framework in the realtime IoT monitoring system to examine its practical use.

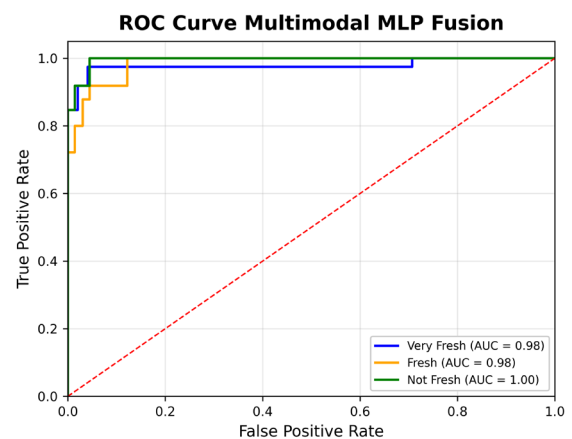


Figure 18. Multiclass ROC Curve of the Multimodal MLP Fusion Model

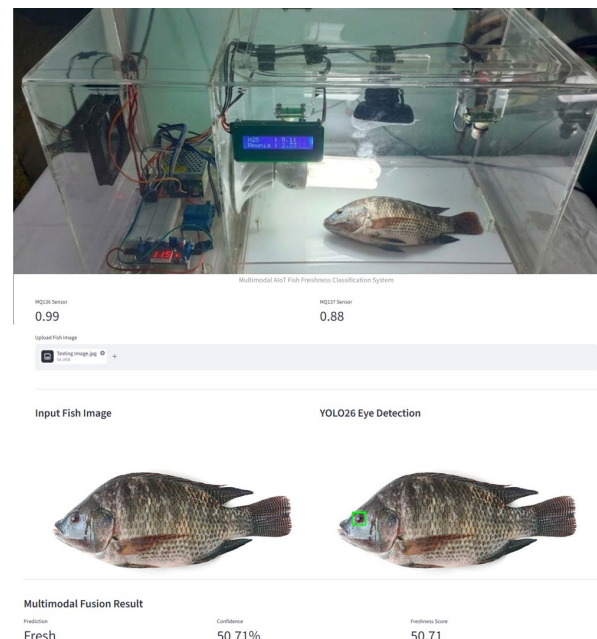


Figure 19. Realtime Deployment of the Proposed Multimodal Fish Freshness System

Figure 19 shows a monitoring dashboard that brings together MQ136 and MQ137 readings, YOLO26 eye localization, and MLP fusion predictions. In the displayed deployment example, the system assigns Fresh with 50.71% confidence while showing the detected eye and current sensor values. This example demonstrates the operation of the integrated workflow for practical IoT freshness assessment.

3.6 Discussion

The discussion examines the framework through comparisons with alternative detection and visual classification models. Table 10 compares YOLO variants evaluated on the same fish image dataset, providing a common basis for assessing eye localization.

Among the models in Table 10, YOLO26 records the highest detection measures: 0.930 precision, 0.896 recall, 0.953 mAP@0.5, and 0.619 mAP@0.5:0.95. Table 11 extends the comparison to visual classification, assessing EfficientNet B3 plus CBAM alongside attention and transformer variants on the same eye dataset.

Table 10. Performance Comparison of YOLO Based Fish Eye Detection Models

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv9	0.871	0.842	0.901	0.548
YOLOv10	0.886	0.857	0.918	0.571
YOLOv11	0.902	0.874	0.927	0.589
YOLOv12	0.914	0.886	0.935	0.603
Research Model (YOLO26)	0.9300	0.896	0.953	0.619

Table 11. Performance Comparison of EfficientNet B3 Based Classification Models

Model	Accuracy	Precision	Recall	F1-Score
EfficientNet-B3	0.86	0.87	0.86	0.87
EfficientNet-B3 + SE Attention	0.88	0.88	0.88	0.88
EfficientNet-B3 + Spatial Attention	0.88	0.89	0.88	0.88
EfficientNet-B3 + ViT	0.89	0.89	0.89	0.89
EfficientNet-B3 + Transformer Attention	0.90	0.90	0.90	0.90
EfficientNet-B3 + CBAM (Research Model)	0.92	0.92	0.92	0.92

The CBAM variant leads Table 11 with 0.92 for accuracy, precision, recall, and F1 score. Its accuracy exceeds standard EfficientNet B3 by approximately 6 percentage points and the Transformer Attention variant by 2 percentage points. These empirical differences support CBAM's contribution to adaptive refinement of the eye features used in freshness classification.

4. Conclusions

The developed framework joins YOLO26 eye localization, EfficientNet B3 plus CBAM visual analysis, BiLSTM temporal sensing, and MLP probability fusion. Detection reaches 0.953 mAP@0.5; the visual and sensor classifiers reach 0.92 and 0.93 accuracy, respectively. Combining their outputs yields the strongest reported classification results, with 0.97 accuracy, 0.97 macro F1, and AUC up to 1.00. Grad CAM highlights eye regions associated with deterioration, supporting interpretation of the visual decisions. Practical deployment still requires systematic

assessment of inference latency, throughput, hardware efficiency, sensor calibration, and environmental compensation. Statistical significance tests and confidence intervals were also outside the present evaluation. Future work should address these measurements to establish the robustness and operational value of the observed improvements.

Declarations

AI Usage Statement

During the manuscript revision process, the authors used ChatGPT (OpenAI) to assist with language editing. This editorial assistance did not alter the experimental data or the reported numerical results. The authors bear full responsibility for the scientific content, interpretation, accuracy, and integrity of the manuscript.

Author Contributions

Rian Ordila: conceptualization, investigation, software, and writing—original draft. Miratu Megasari: data curation, validation, and writing—review and editing. Yulanda: investigation, data curation, and visualization. Yuda Irawan: methodology, formal analysis, supervision, and writing—review and editing. Refni Wahyuni: validation, supervision, and writing—review and editing. All authors reviewed and approved the final manuscript.

Funding Statement

This research is funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology with the Regular Fundamental Research Scheme in 2026.

Conflict of Interest Statement

The authors declare no financial, professional, or personal conflicts of interest that could have influenced the research or its reporting.

Ethical Clearance

The study assessed harvested fish through image analysis and spoilage-related gas measurements. It did not involve human participants or experimental procedures on live animals. Sample handling, data collection, and analysis were conducted in accordance with applicable institutional ethical standards.

Data and Code Availability

The visual datasets were obtained from Roboflow. The processed sensor data and implementation code supporting this study are available from the corresponding author upon reasonable request, subject to applicable permissions and dataset redistribution conditions.

References

- [1] A. D. Cortés-Sánchez *et al.*, "Processing, Quality and

- Elemental Safety of Fish,” 2024. doi: [10.3390/app14072903](https://doi.org/10.3390/app14072903).
- [2] D.-Y. Kim, S.-W. Park, and H.-S. Shin, “Fish Freshness Indicator for Sensing Fish Quality during Storage,” *Foods*, vol. 12, no. 9, 2023, doi: [10.3390/foods12091801](https://doi.org/10.3390/foods12091801).
- [3] A.-F. M. El-Sayed and K. Fitzsimmons, “From Africa to the world—The journey of Nile tilapia,” *Rev. Aquac.*, vol. 15, no. S1, pp. 6–21, Feb. 2023, doi: <https://doi.org/10.1111/raq.12738>.
- [4] M. E. Abd El-Hack et al., “Effect of environmental factors on growth performance of Nile tilapia (*Oreochromis niloticus*),” *Int. J. Biometeorol.*, vol. 66, no. 11, pp. 2183–2194, 2022, doi: [10.1007/s00484-022-02347-6](https://doi.org/10.1007/s00484-022-02347-6).
- [5] D. Shen, M. Zhang, A. S. Mujumdar, and Y. Ma, “Consumer-oriented smart dynamic detection of fresh food quality: recent advances and future prospects,” *Crit. Rev. Food Sci. Nutr.*, vol. 64, no. 30, pp. 11281–11301, Nov. 2024, doi: [10.1080/10408398.2023.2235703](https://doi.org/10.1080/10408398.2023.2235703).
- [6] M. N. Madhubhashini, C. P. Liyanage, A. U. Alahakoon, and R. P. Liyanage, “Current applications and future trends of artificial senses in fish freshness determination: A review,” *J. Food Sci.*, vol. 89, no. 1, pp. 33–50, Jan. 2024, doi: <https://doi.org/10.1111/1750-3841.16865>.
- [7] M. Kundu, A. Vashisth, and P. Krishnan, “Instrumental detection of fish freshness in research and food industry: a review,” *J. Food Sci. Technol.*, vol. 63, no. 1, pp. 22–50, 2026, doi: [10.1007/s13197-025-06497-4](https://doi.org/10.1007/s13197-025-06497-4).
- [8] D. Li and L. Du, “Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish,” *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 4077–4116, 2022, doi: [10.1007/s10462-021-10102-3](https://doi.org/10.1007/s10462-021-10102-3).
- [9] R. Singh, C. Nickhil, R. Nisha, K. Upendar, B. Jithender, and S. C. Deka, “A Comprehensive Review of Advanced Deep Learning Approaches for Food Freshness Detection,” *Food Eng. Rev.*, vol. 17, no. 1, pp. 127–160, 2025, doi: [10.1007/s12393-024-09385-3](https://doi.org/10.1007/s12393-024-09385-3).
- [10] S. U. Saba et al., “A Critical Review on the Application and Innovation in Smart Fisheries,” *Eng. Reports*, vol. 8, no. 1, p. e70608, Jan. 2026, doi: <https://doi.org/10.1002/eng2.70608>.
- [11] S. M. Hamidy, Y. Kuvvetli, Y. Sakarya, S. T. Özkütük, and Y. Özoğlu, “Deep Learning-Based Prediction of Fish Freshness and Purchasability Using Multi-Angle Image Data,” 2026. doi: [10.3390/foods15010068](https://doi.org/10.3390/foods15010068).
- [12] S. Sattar, T. Abbas, M. Tabish, and G. Zhiqiang, “Computer vision in aquaculture: transforming fish freshness monitoring,” *Crit. Rev. Food Sci. Nutr.*, pp. 1–29, Dec. 2025, doi: [10.1080/10408398.2025.2607533](https://doi.org/10.1080/10408398.2025.2607533).
- [13] M. Ashimgaliyev, E. R. Joseph, A. Zhumadillayeva, and A. Baimakhanova, “Development of Deep Learning Methods for Visual Document Classification Using Hybrid Vision Transformer–EfficientNet Architecture,” *IEEE Access*, vol. 14, pp. 28041–28053, 2026, doi: [10.1109/ACCESS.2025.3646481](https://doi.org/10.1109/ACCESS.2025.3646481).
- [14] M. Hayat, N. Ahmad, A. Nasir, and Z. A. Tariq, “Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification,” *IEEE Access*, vol. 12, pp. 184119–184131, 2024, doi: [10.1109/ACCESS.2024.3503413](https://doi.org/10.1109/ACCESS.2024.3503413).
- [15] K. M. A. Ramadan, E. S. A. Bendary, H. B. Khalil, S. A. Ali, A. R. Ahmed, and M. A. A. Mahmoud, “Smart Detection of Food Spoilage Using Microbial Volatile Compounds: Technologies, Challenges, and Future Outlook,” *J. Agric. Food Chem.*, vol. 73, no. 31, pp. 19222–19243, Aug. 2025, doi: [10.1021/acs.jafc.5c05927](https://doi.org/10.1021/acs.jafc.5c05927).
- [16] M. Yumnam, K. Gopalakrishnan, S. Dhua, Y. Srivastava, and P. Mishra, “A Comprehensive Review on Smartphone-Based Sensor for Fish Spoilage Analysis: Applications and Limitations,” *Food Bioprocess Technol.*, vol. 17, no. 12, pp. 4575–4597, 2024, doi: [10.1007/s11947-024-03391-3](https://doi.org/10.1007/s11947-024-03391-3).
- [17] H. G. Nanditha, S. Nishita, G. R. Kavana, G. Lakshmikanth, Y. Venkat, and K. Bandyopadhyay, “Innovative Wireless Technology for Monitoring Consumable Product Degradation,” in *Emerging Electronics and Automation*, M. Kankanhalli, S. Bhartiya, and P. S. Pravin, Eds., Singapore: Springer Nature Singapore, 2026, pp. 227–242, doi: https://doi.org/10.1007/978-981-96-9554-6_19.
- [18] L. Zhang, G. Ren, J. Du, S. Li, Y. Li, and D. Xu, “A novel spatio-temporal attention-based bidirectional LSTM model for moisture content prediction in drying process,” *Dry. Technol.*, vol. 42, no. 14, pp. 2122–2136, Nov. 2024, doi: [10.1080/07373937.2024.2406441](https://doi.org/10.1080/07373937.2024.2406441).
- [19] V. Viancy and N. Gobalakrishnan, “Intelligent food quality monitoring: A hybrid dilated convolutional network and BiLSTM-MSVM model for robust beef quality assessment,” *Food Control*, vol. 176, p. 111396, 2025, doi: <https://doi.org/10.1016/j.foodcont.2025.111396>.
- [20] P. Padmanabha et al., “Precision Without Complexity: A Comparative Study of YOLOv2 Pose Variants for Distal Arm Landmark Detection,” 2026. doi: [10.3390/app16083968](https://doi.org/10.3390/app16083968).
- [21] B. Çarklı Yavuz, “Scale-Dependent Performance Analysis of YOLOv2 and YOLOv11 for PPE Detection,” 2026. doi: [10.3390/electronics15061146](https://doi.org/10.3390/electronics15061146).
- [22] A. A. Abd El-Aziz, M. A. Mahmood, and S. A. El-Ghany, “EfficientNet-B3-Based Automated Deep Learning Framework for Multiclass Endoscopic Bladder Tissue Classification,” *Diagnostics*, vol. 15, no. 19, 2025, doi: [10.3390/diagnostics15192515](https://doi.org/10.3390/diagnostics15192515).
- [23] B. Chen and Z. Dang, “Fast PCB Defect Detection Method Based on FasterNet Backbone Network and CBAM Attention Mechanism Integrated With Feature Fusion Module in Improved YOLOv7,” *IEEE Access*, vol. 11, pp. 95092–95103, 2023, doi: [10.1109/ACCESS.2023.3311260](https://doi.org/10.1109/ACCESS.2023.3311260).
- [24] U. G. Redwan, T. Zaman, and H. Bin Mizan, “Spatio-temporal CNN-BiLSTM dynamic approach to emotion recognition based on EEG signal,” *Comput. Biol. Med.*, vol. 192, p. 110277, 2025, doi: <https://doi.org/10.1016/j.combiomed.2025.110277>.
- [25] Muhardi, M. R. Amarthia, R. Melyanti, and Y. Irawan, “Optimization of a New Adaptive Stacking Ensemble Model Integrated with IoT for Stress Level Detection Based on Physiological Signals,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 10, no. 1 SE-Computer Science Applications, Feb. 2026, doi: [10.29207/resti.v10i1.6770](https://doi.org/10.29207/resti.v10i1.6770).
- [26] H. Fonda, Yulanda, M. Ikhsanudin, Muhardi, and Y. Irawan, “Application of Certainty Factor Method to Identify Pests in Crystal Jamboo Plants Application of Certainty Factor Method to Identify Pests in,” *J. Phys. Conf. Ser.*, vol. 1783, no. 1, p. 012053, 2021, doi: [10.1088/1742-6596/1783/1/012053](https://doi.org/10.1088/1742-6596/1783/1/012053).
- [27] Y. Irawan, S. Defit, and R. Sovia, “HSEL: An adaptive rotational hybrid stacking ensemble with exponential accuracy weighting for real-time forest fire risk prediction,” *Eng. Appl. Sci. Res.*, vol. 53, no. 5 SE-ORIGINAL RESEARCH, pp. 519–533, Sep. 2026, doi: [10.64960/easr.2026.266473](https://doi.org/10.64960/easr.2026.266473).