E-ISSN: 2963-4946
Vol. 1 No. 4 January 2023, pp. 192-209
DOI: 10.59888/ajosh.v4i10.736

Asian Journal of Social and Humanities

Fine-Tuning Strategies of EfficientNetV2B0 for Fish Freshness Classification Based on Eye Images with Grad-CAM

Tio Bagus Purnama Apriliano
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon, Indonesia
Corresponding Author: tiopurnamaapriliano@gmail.com

Keywords:
fish freshness; EfficientNetV2B0; fine-tuning; transfer learning; Grad-CAM

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author.
Published by
Pelopor Publikaasi Akademika

Abstract

Background: Using eye images to classify fish freshness requires an adaptation configuration of a pretrained model that suits the target data, yet the effect of fine-tuning depth on compact backbones remains rarely examined.
Objective: This study analysed four training configurations of EfficientNetV2B0 and interpreted their predictions using Grad-CAM.
Method: Version 2 of the Multistage Fish Eyes Dataset contains 4,800 images in three classes (Fresh, Highly Fresh, Not Fresh), partitioned into 3,360 training, 720 validation, and 720 test images. The configurations were a frozen backbone and the unfreezing of approximately the last 20%, the last 50%, and the whole backbone, each with ImageNet weights, 224 × 224-pixel inputs, and one seed.
Results: Test accuracy reached 82.36%, 96.67%, 97.22%, and 87.92%, respectively. Unfreezing approximately the last 50% of layers produced the lowest validation loss and the highest macro F1 (97.23%), with only four more correct predictions than the 20% configuration, a non-significant difference (exact McNemar $p \geq 0.125$). Grad-CAM overlays showed attribution on the eye and surrounding head region but did not include the Not Fresh class.
Conclusion: Partial fine-tuning is a promising and computationally economical option on the partition used, although the findings remain exploratory because one seed was used, training factors changed together, and fish identity separation was not audited.

INTRODUCTION

Freshness is a quality attribute of fish that changes during storage and distribution. These changes can be observed through the appearance of the eyes, gills, and other physical characteristics. In conventional practice, sensory schemes such as the Quality Index Method score several attributes, including the eyes, gills, skin, and odour, which makes the assessment dependent on trained panellists and difficult to scale (Franceschelli et al., 2021). Sensing and image-processing technologies open opportunities to support more measurable assessment.

Franceschelli et al., (2021) reviewed a range of technologies for assessing fish freshness and safety and showed that measurement targets need to be distinguished according to the characteristics of each indicator. Reliable freshness assessment is also

important from an economic and public-health perspective. Global fisheries and aquaculture production reached a record 223.2 million tonnes in 2022, of which 185.4 million tonnes were aquatic animals, and apparent consumption of aquatic animal foods rose to 20.7 kg per capita (FAO, 2024).

Indonesia is among the main producers of aquatic animals, accounting for approximately 7% of global production (FAO, 2024), so the quality of fish reaching domestic markets affects a large number of consumers and producers. At the same time, approximately 35% of the global harvest is estimated to be lost or wasted between capture and consumption, partly because of spoilage during handling, storage, and distribution (FAO, 2020). Because fish deteriorate rapidly after harvest, inaccurate freshness judgement can cause economic losses and allow products of questionable quality to be sold, particularly in traditional markets where inspection is informal. Rapid, objective, and low-cost screening methods are therefore needed to complement conventional quality control.

Eye images are one of the inputs that have been studied for recognizing freshness levels. Prasetyo et al. (2022) developed a modified MobileNetV1 for classifying eye images, whereas Akgül et al., (2023) combined detection and transfer learning on eye and gill characteristics. Yasin et al. (2023) used features from pretrained networks together with machine-learning classifiers on fish-body images labelled as fresh or stale. Mohammadi L et al. (2020) also combined eye and gill colour features with machine-learning classifiers for fish freshness classification. These studies support the feasibility of image-based quality recognition, but differences in datasets, input modalities, numbers of classes, and evaluation protocols limit direct comparison of performance figures.

Most image-based studies on fish-eye freshness have used the Freshness of the Fish Eyes dataset (Prasetyo et al., 2022). MFED was released only recently, and at the time of this revision no peer-reviewed study reporting classification results on MFED version 1 or version 2 could be identified. The design of MFED, which photographs the same refrigerated fish repeatedly under controlled lighting, background, and angle settings across six storage days, differs from that of earlier datasets and introduces a specific evaluation risk, namely the presence of images of the same individual in several partitions (Kapoor & Narayanan, 2023). Results on MFED therefore cannot yet be benchmarked against prior work on the same data and should be regarded as an initial reference point for this dataset.

Table 1 summarizes these image-based studies. The reported accuracies range widely, from 63.21% on the eight-species Freshness of the Fish Eyes dataset to 100% for a two-class fish-body image dataset, which illustrates how strongly the figures depend on the number of species and classes, the body region analysed, and the evaluation protocol. A higher number in one study therefore cannot be read as evidence that its method would outperform another method on a different dataset.

Table 1. Summary of related image-based fish freshness studies

Study	Data and classes	Method	Reported result
Prasetyo et al. (2022)	Freshness of the Fish Eyes dataset; eye images of eight species, three freshness levels per species	MobileNetV1 bottleneck with expansion (MB-BE)	Accuracy 63.21%
Akgül et al., (2023)	Anchovy and horse mackerel; fresh vs. non-fresh; eye and gill regions	YOLOv5 detection + Inception-ResNet-v2 or Xception	Accuracy up to 97.67% (anchovy) and 96.0% (horse mackerel)
Yasin et al. (2023)	4,476 fish-body images; fresh vs. stale	SqueezeNet/InceptionV 3 features + SVM, ANN, LR	Accuracy 100% (SVM, ANN, and LR)
Mohammadi Lalabadi et al. (2020)	Rainbow trout over 10 days of ice storage; colour features of eyes and gills	RGB, HSV, and L*a*b* colour features + multi-class ANN and SVM	ANN accuracy about 84% (eyes) and 96% (gills)
This study	MFED version 2; 4,800 eye images of two species; three classes	EfficientNetV2B0 with four adaptation configurations	Test accuracy 82.36–97.22%

Source: Data Processed

Transfer learning allows a model to exploit representations learned from pretraining data by freezing or updating weights. The choice of which layers to unfreeze is part of the adaptation configuration, together with the learning rate and the handling of normalization. Kim et al., (2022) discussed these adaptation choices in a review of medical image classification; their relevance to the present study is used as a methodological basis, not as evidence of results on fish. EfficientNetV2 was chosen as a model family designed with attention to training efficiency (Tan & Le, 2021).

In practice, adaptation strategies form a spectrum. At one end, the pretrained backbone is used as a fixed feature extractor and only a new classification head is trained; at the other end, all backbone weights are updated with a small learning rate. Intermediate strategies unfreeze only the deeper layers, which encode more task-specific features, while the earlier layers that capture generic edges and textures remain fixed. The appropriate point on this spectrum depends on the size of the target dataset and its similarity to the pretraining data (Kim et al., 2022). Because close-up fish-eye images differ considerably from the natural images in ImageNet, it is not obvious in advance whether a frozen, partial, or full adaptation will perform best.

Evidence from general computer vision supports this reasoning. Yosinski et al. (2014) showed that features in the early layers of convolutional networks are largely general, whereas those in the deeper layers become increasingly specific to the pretraining task, so that transferability decreases as the source and target tasks become more dissimilar. Kornblith et al. (2019) found that ImageNet accuracy is strongly correlated

with transfer accuracy, yet on some fine-grained datasets the benefit of ImageNet pretraining was small, indicating that pretrained features do not transfer uniformly across target domains.

The choice of fine-tuning depth also has practical consequences. Updating more layers increases the number of trainable parameters, the memory required for gradients and optimizer states, and the training time. These costs matter for small laboratories, for repeated retraining when new species or acquisition conditions are added, and for screening tools intended to run on mobile devices (Tan & Le, 2021). Identifying the shallowest adaptation that achieves acceptable accuracy is therefore relevant beyond a single benchmark.

The gap examined in this study is the need for empirical evidence on the differences between four EfficientNetV2B0 configurations on a partition of version 2 of the Multistage Fish Eyes Dataset (MFED). The previous studies cited above do not directly answer this question for this set of configurations and this partition. The contribution of the study is limited to an internal comparison of configurations, an analysis of errors between classes, and a visual interpretation of predictions.

This study does not propose a new architecture and does not claim that fine-tuning and Grad-CAM have not been used before.Within these limits, the novelty of the study lies in three elements. It provides, to the best of the author’s knowledge, one of the first reported comparisons of adaptation depths on MFED version 2; it relates overall metrics to the dominant confusion between adjacent freshness levels; and it reports the number of trainable parameters of each configuration, so that accuracy can be weighed against the computational resources required for deployment.

The aim of the study is to analyse the performance of the four configurations, examine the relationship between validation and test results, and describe the location of attributions in the predictions of the selected configuration. The evaluation distinguishes checkpoint selection through validation from the selection between configurations, which in the actual implementation used test results. This distinction is necessary because using test data for selection decisions limits the independence of the evaluation (Kapoor & Narayanan, 2023).

Grad-CAM is used as a supplementary interpretation, bearing in mind that a plausible-looking saliency map does not prove that the model's explanation is trustworthy (Arun et al., 2021). To limit this dependence, a validation-only selection rule is additionally applied, under which the configuration with the lowest validation loss is identified without reference to the test results. Accordingly, the study addresses three questions: first, how do the four adaptation configurations of EfficientNetV2B0 differ in overall and per-class test performance on MFED version 2; second, is the ranking of the configurations based on validation loss and validation accuracy consistent with their ranking on the test partition; and third, which image regions are emphasized by Grad-CAM in the predictions of the selected configuration, and what can and cannot be concluded from these maps.

METHOD

Data and Research Design

This study used a quantitative comparative experiment on a single architecture with four training configurations. The overall workflow, from data loading to visual interpretation, is shown in Figure 1. The data source was MFED version 2, published on Mendeley Data (https://doi.org/10.17632/67nmx3mhwh.2) and licensed under CC BY 4.0. The dataset consists of 4,800 eye images of two freshwater fish species, *Cyprinus carpio* var. Fanli and *Carassius auratus*. The fish were stored at 4 °C for six days and photographed under white and warm yellow light, on a chopping board and on A4 paper backgrounds, on both sides of the eye, and from several angles.

The source labels are Highly Fresh for storage of up to two days, Fresh for more than two and up to four days, and Not Fresh for more than four days. Each class in the source contains 1,600 images. The labels were used as provided by the contributor, without reassignment through chemical or microbiological examination. This characterization limits the model output to the classification of freshness levels as defined by the dataset.

Table 2. Image partitions loaded in the experiment

Partition	Number of images	Proportion
Training	3,360	70%
Validation	720	15%
Test	720	15%
Total	4,800	100%

Source: Data Processed

The notebook loaded partitions that were already available in a Google Drive directory (Table 2). The test data contained 240 images per class, whereas the class distributions of the training and validation sets were not printed. The process used to create the partitions, any duplicate audit, and the separation of fish identities were not documented. Because individual fish were photographed repeatedly, fish identity is an important unit for leakage audits; the results do not claim that independence between partitions has been demonstrated (Kapoor & Narayanan, 2023).

Two characteristics of the partitions can nevertheless be derived from the available information. First, because each source class contains 1,600 images and the test partition contains 240 images per class, the combined training and validation partitions necessarily contain 1,360 images per class; a 70:15 ratio applied within each class would correspond to 1,120 training and 240 validation images per class, although this stratification could not be verified from the notebook output. Second, according to the dataset description, each fish was photographed 40 times per storage day (both eyes, five angles, and four lighting–background combinations), so that the 4,800 images correspond to 120 fish-days, or approximately 20 individuals if each fish was photographed on every day.

With so few individuals, an image-level random split is very likely to place images of the same fish, and of the same fish on the same day, in all three partitions. A re-split by fish identity followed by re-training of all configurations could not be completed for

this revision; the absolute accuracies are therefore interpreted as performance on seen individuals under controlled acquisition, not as generalisation to new fish.

Preprocessing and Architecture

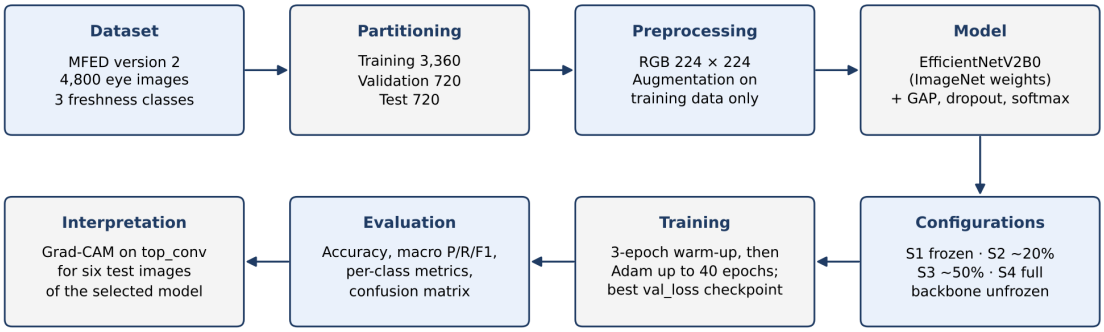


Figure 1. Research workflow

Images were loaded as RGB images of 224 × 224 pixels with a batch size of 32. EfficientNetV2B0 used ImageNet weights, without its built-in classifier, and with include_preprocessing=True. Inputs were in the pixel range 0–255 and were therefore not divided by 255 before internal scaling. These settings are the implementation configuration of this experiment. Training augmentation consisted of horizontal RandomFlip, RandomRotation 0.04, RandomZoom 0.10, RandomTranslation 0.05 on both axes, and RandomContrast 0.10. Validation and test data received no random augmentation. Augmentation was used to increase the variation of training inputs while considering the suitability of each transformation for the classification task (Mumuni & Mumuni, 2022).

EfficientNetV2 combines Fused-MBConv blocks in the early stages, where a single regular 3 × 3 convolution replaces the expansion and depthwise convolutions of MBConv, with MBConv blocks and squeeze-and-excitation in the later stages; the family was obtained through training-aware neural architecture search and scaling (Tan & Le, 2021). B0 is the smallest baseline member of the family. The backbone output was passed to Global Average Pooling, dropout 0.30, and a three-unit Dense layer with softmax. Computation used mixed_float16 with a float32 output layer, together with caching and prefetching for data loading.

Configurations and Training

Each scenario built a new model and ran a three-epoch warm-up with a frozen backbone and a learning rate of 0.001. The main stage used Adam and sparse categorical cross-entropy for a maximum of 40 epochs. The unfreezing percentages in Table 3 were calculated from the indices of backbone.layers, not from the proportion of parameters. Rounding of the boundary and the exclusion of Batch Normalization layers make these percentages approximate.

Table 3. Training configurations in the main stage

Code	Backbone layers trained	Learning rate	Batch Normalization
S1	None	0.001	Frozen
S2	Approximately the last 20%, excluding BN	0.0001	Frozen
S3	Approximately the last 50%, excluding BN	0.0001	Frozen
S4	Entire backbone	0.00001	Trained

Source: Data Processed

EarlyStopping monitored val_loss with a patience of 7 and restored the best weights. ReduceLROnPlateau used the same monitor, a factor of 0.5, a patience of 3, and a minimum learning rate of 0.0000001. ModelCheckpoint saved the checkpoint with the best val_loss, and seed 42 was applied to Python, NumPy, and TensorFlow. The experiment was run once per configuration on Google Colab with T4 GPU metadata; the full dependency versions and actual durations were not recorded. Because the unfreezing percentages refer to layer indices rather than parameters, the number of trainable parameters was reconstructed from the Keras implementation of EfficientNetV2B0 (TensorFlow 2.21; 270 backbone layers) using the same unfreezing rule, with the moving statistics of Batch Normalization counted as non-trainable.

Including the 3,843 parameters of the classification head, the trainable parameters numbered 3,843 for S1, approximately 2.22 million for S2, 4.57 million for S3, and 5.86 million for S4. Unfreezing approximately 20% and 50% of the layers therefore corresponds to approximately 38% and 78% of the trainable backbone parameters, respectively, because the deeper MBConv blocks are the widest. The counts were identical whether the layer boundary was truncated or rounded. Training duration cannot be reconstructed in this way and is not reported. The depth of unfreezing changed together with the learning rate and, in S4, also with the status of Batch Normalization.

The comparison therefore measures configuration packages, not the isolated causal effect of fine-tuning depth. Each model also ran its own warm-up, with no evidence that a single identical initial classification-head checkpoint was used. These limitations are retained in the interpretation of the results.Keeping Batch Normalization layers frozen in S2 and S3 means that their running mean and variance remain those estimated on ImageNet, while only the convolutional weights of the unfrozen blocks are updated.

In S4, the normalization statistics were also updated using the fish-eye images. With small batches and a domain that differs from the pretraining data, updating these statistics can change the activations of the whole network at the start of fine-tuning, which is one reason why frozen normalization is often recommended for partial adaptation. The present design cannot separate this factor from the lower learning rate used in S4.

Evaluation and Interpretation

Evaluation covered accuracy, macro precision, macro recall, macro F1, per-class metrics, and the confusion matrix. Reporting several metrics is necessary to distinguish overall performance from the error pattern of each class (Hicks et al., 2022). For a class c, precision, recall, and F1 were computed as in Equations (1)–(3), where TP, FP, and FN are the true positives, false positives, and false negatives of that class.

$$\text{Precision}_c = TP_c / (TP_c + FP_c) \quad (1)$$

$$\text{Recall}_c = TP_c / (TP_c + FN_c) \quad (2)$$

$$F1_c = 2 \times \text{Precision}_c \times \text{Recall}_c / (\text{Precision}_c + \text{Recall}_c) \quad (3)$$

Macro F1 is the average F1 of the three classes, as in Equation (4) with K = 3, whereas accuracy is the proportion of correct predictions among all images. Because the test set is balanced, with 240 images per class, macro recall is numerically equal to accuracy in this experiment; macro precision and macro F1 can still differ when errors are concentrated in particular predicted classes.

$$\text{Macro F1} = (1/K) \Sigma_c F1_c \quad (4)$$

The checkpoint of each scenario was selected based on val_loss, but the actual code selected the configuration for Grad-CAM based on the highest test macro F1. The comparison of minimum val_loss between configurations was carried out retrospectively from the available logs. The test results are therefore reported as exploratory rather than as an independent confirmatory estimate of a configuration fixed in advance. No significance testing or estimation of variation across seeds was performed. To address the use of test results in the original selection, a validation-only rule was applied retrospectively to the logged checkpoints: the configuration with the lowest minimum val_loss, with validation accuracy as a secondary criterion, is selected before any test result is consulted.

Under this rule S3 is selected by both criteria (Table 4), which is the same configuration identified by test macro F1; the test results of S3 can therefore be read as the evaluation of a configuration that validation alone would have chosen, although the test set had been inspected during the original analysis. Uncertainty in test accuracy is expressed with 95% Wilson confidence intervals for a binomial proportion on 720 images. For the paired comparison of S3 and S2, McNemar's test (Dietterich, 1998; McNemar, 1947) depends only on the discordant images, namely those classified correctly by one configuration and incorrectly by the other.

Although the per-image predictions were not exported, the difference of four correct predictions implies that the discordant counts satisfy $b - c = 4$, where b and c are the numbers of images correct only for S3 and only for S2. The exact two-sided p-value is smallest when $b = 4$ and $c = 0$ and increases with every additional discordant pair, so this case provides a lower bound for the p-value of the comparison. Grad-CAM used the top_conv layer with gradients with respect to the softmax probability of the predicted

class. The gradients were averaged to weight the activations, followed by ReLU, normalization, and an overlay with the JET colour map at an alpha of 0.4.

Six images were selected using unbatch, a shuffle with a buffer of 500 and seed 42, and take(6), without stratification by class or prediction status. For a feature map A^k of top_conv and the score y^c of the predicted class c, the channel weights and the class activation map are computed as in Equations (5) and (6), where Z is the number of spatial positions; the map is then upsampled to the input size before the overlay.

$$\alpha_k^c = (1/Z) \Sigma_i \Sigma_j \partial y^c / \partial A_{ij}^k \quad (5)$$

$$L_{Grad\text{-}CAM}^c = ReLU(\Sigma_k \alpha_k^c A^k) \quad (6)$$

The use of the softmax probability rather than the pre-softmax class score departs from the original formulation of Grad-CAM, in which the score of class c is the logit (Selvaraju et al., 2020). The gradient of the softmax probability equals that probability multiplied by the difference between the gradient of the class logit and the probability-weighted average of the gradients of all logits; as the predicted probability approaches one, this difference approaches zero, so gradients can saturate for confident predictions and become influenced by competing classes.

Normalisation of the map to the range 0–1 removes differences in overall magnitude but not changes in spatial pattern caused by competing classes. The maps in this study are therefore treated as qualitative indications only, and recomputation with the logit of the predicted class is recommended for follow-up work.The interpretation follows the available outputs and is not treated as segmentation validation or as biological evidence about the model (Arun et al., 2021).

RESULTS AND DISCUSSION

Training and Validation Results

Table 4. Best checkpoints based on the minimum validation loss in the main stage

Code	Epochs run	Best epoch	Val loss	Train acc (%)	Val acc (%)
S1	40	40	0.4889	77.02	81.81
S2	40	33	0.0647	98.45	97.50
S3	37	30	0.0304	99.82	98.89
S4	40	36	0.3128	97.44	86.81

Source: Data Processed

S3 reached the lowest minimum validation loss, namely 0.0304 in main-stage epoch 30, with a validation accuracy of 98.89% (Table 4). S1, S2, and S4 completed 40 main-stage epochs, whereas S3 stopped at epoch 37. These numbers do not include the three warm-up epochs of each scenario. The checkpoint figures come from a reconstruction of the logs rounded to four decimals, without re-verification of the weight files.

S1 reached its best validation loss in the final epoch, which suggests that the frozen-backbone configuration had not converged within the 40-epoch budget. Its training

accuracy of 77.02% was lower than its validation accuracy of 81.81%, a pattern that can occur when dropout and augmentation are active during training but not during validation. Together, these observations indicate that ImageNet features alone, without adaptation of the deeper blocks, were insufficient to separate the three freshness classes well in this configuration.

S4 showed a training accuracy of 97.44% and a validation accuracy of 86.81% at its best checkpoint, a gap of approximately 10.63 percentage points. This pattern indicates a mismatch in performance between the two partitions for this configuration. Its cause cannot yet be separated between the unfreezing of layers, the learning rate, the updating of Batch Normalization, and the state of optimization. The results are not sufficient to conclude that fine-tuning the entire backbone is always worse.

Table 5 compares the validation accuracy at the selected checkpoint with the test accuracy of each configuration. The ordering of the four configurations was identical on both partitions, with S3 first, S2 second, S4 third, and S1 last, and the absolute differences between validation and test accuracy were below 1.7 percentage points. This agreement suggests that the validation set was informative for ranking the configurations in this experiment. It does not, however, show that the partitions are independent of each other, because images of the same fish could be present in more than one partition.

Table 5. Validation accuracy at the selected checkpoint compared with test accuracy

Code	Val acc (%)	Test acc (%)	Test − val (pp)	Rank (val / test)
S1	81.81	82.36	+0.55	4 / 4
S2	97.50	96.67	−0.83	2 / 2
S3	98.89	97.22	−1.67	1 / 1
S4	86.81	87.92	+1.11	3 / 3

Source: Data Processed

Test Performance and Errors between Classes

Table 6. Performance of the four configurations on 720 test images

Code	Accuracy (%)	Macro precision (%)	Macro recall (%)	Macro F1 (%)	Errors
S1	82.36	82.23	82.36	82.21	127
S2	96.67	96.75	96.67	96.68	24
S3	97.22	97.28	97.22	97.23	20
S4	87.92	88.60	87.92	88.04	87

Source: Data Processed

S3 produced 700 correct and 20 incorrect predictions, whereas S2 produced 696 correct and 24 incorrect predictions (Table 6). The difference in accuracy between the two configurations is about 0.56 percentage points, or four images. S1 and S4 produced 593 and 633 correct predictions, respectively. The ordering of the results supports partial

fine-tuning as a candidate in this experiment, but a single run does not show that the ordering would remain stable in other repetitions. Expressed with 95% Wilson confidence intervals, test accuracy was 82.36% (79.41–84.97%) for S1, 96.67% (95.09–97.75%) for S2, 97.22% (95.75–98.19%) for S3, and 87.92% (85.33–90.10%) for S4; the intervals of S2 and S3 overlap almost entirely.

Table 7. Per-class metrics of configurations S2 and S3

Code	Class	Precision (%)	Recall (%)	F1 (%)
S2	Fresh	93.23	97.50	95.32
S2	Highly Fresh	99.15	97.08	98.11
S2	Not Fresh	97.86	95.42	96.62
S3	Fresh	94.35	97.50	95.90
S3	Highly Fresh	98.35	99.17	98.76
S3	Not Fresh	99.13	95.00	97.02

Source: Data Processed

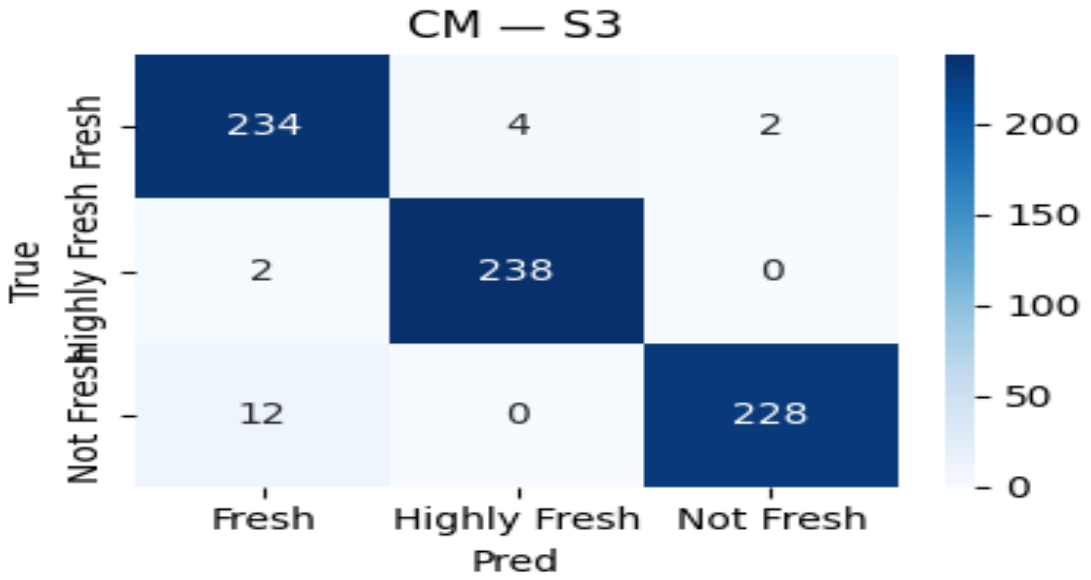


Figure 2. Confusion matrix of S3

The confusion matrix of S3 (Figure 2) shows that 234 Fresh, 238 Highly Fresh, and 228 Not Fresh images were classified correctly. The largest error was 12 Not Fresh images predicted as Fresh, or 60% of all S3 errors. The Not Fresh recall of S3 was 95.00%, slightly lower than the 95.42% of S2 (Table 7). The macro F1 advantage of S3 therefore does not apply uniformly to every class-recognition target.

The remaining errors of S3 were also concentrated between adjacent freshness levels: four Fresh images were predicted as Highly Fresh, two Fresh images as Not Fresh, and two Highly Fresh images as Fresh, while no Highly Fresh image was predicted as Not Fresh or vice versa. This pattern is consistent with the ordinal nature of the labels, which are defined by consecutive storage intervals; visual differences are likely to be smallest

near the boundaries between neighbouring intervals. A treatment of the task as ordinal classification, or the use of storage time as a continuous target, could therefore be considered in future work.

This error pattern needs attention if the intended use emphasizes the recognition of quality deterioration. Selection based on a single overall metric does not necessarily meet the needs of each class, so Not Fresh recall should be reported together with macro F1. Freshness in these results still refers to the MFED labels. Image predictions are not used to establish microbiological safety or fitness for consumption.

Comparison of the Adaptation Configurations

The gain from unfreezing approximately the last 20% of the backbone (S2) rather than keeping it frozen (S1) was large, at 14.31 percentage points of accuracy, whereas the further gain from unfreezing approximately 50% (S3) was only 0.56 percentage points. To put the second difference in perspective, the binomial standard error of an accuracy of about 97% estimated on 720 images is approximately 0.64 percentage points. The difference between S2 and S3 is therefore smaller than one standard error of either estimate. A paired test such as McNemar's test would be more appropriate for two models evaluated on the same images, but it requires per-image predictions, which were not stored.

The ranking of S3 above S2 should thus be regarded as tentative. Even under the assumption most favourable to S3, namely that all discordant images were classified correctly by S3 and incorrectly by S2 ($b = 4$, $c = 0$), the exact McNemar test gives $p = 0.125$, and any additional discordant pairs would increase this value. The difference between S3 and S2 is therefore not statistically significant at $\alpha = 0.05$ under any possible pattern of per-image predictions. In practical terms, S2 reached a statistically indistinguishable level of performance with fewer than half of the trainable parameters of S3 (2.22 million versus 4.57 million), which makes it an attractive option when computational resources are limited.

S4, which updated the whole backbone, performed clearly worse than the partial configurations, with 87.92% test accuracy. Several explanations are compatible with the logs. The learning rate of 0.00001 was ten times smaller than in S2 and S3, so the deeper blocks may have adapted more slowly within the epoch budget; updating Batch Normalization statistics with batches of 32 images may have shifted the pretrained activations; and updating all layers increases the capacity to fit the training partition, which is consistent with the gap between training and validation accuracy in Table 4. Because these factors changed together, the present results do not indicate which of them was decisive. They do suggest that, for this dataset size, adapting only the deeper part of the backbone while keeping normalization fixed was the more robust package among those tested.

Grad-CAM Interpretation

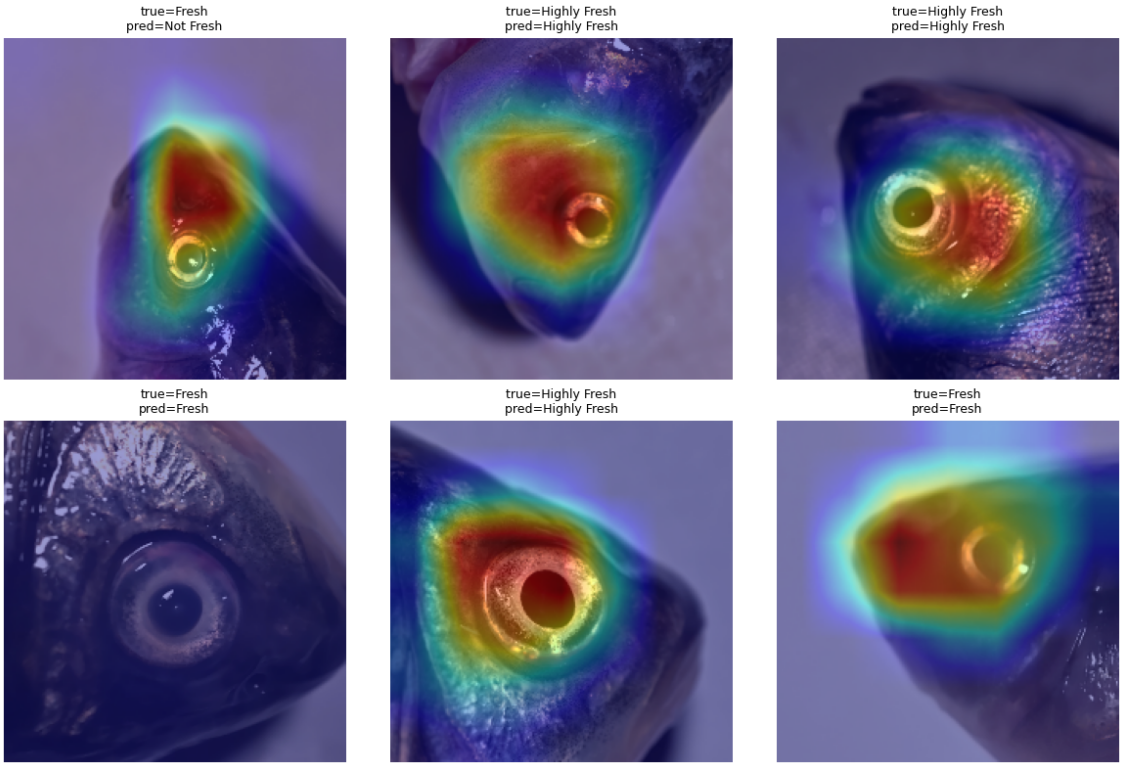


Figure 3. Six Grad-CAM overlays of S3

Five of the six Grad-CAM samples were predicted correctly, whereas one Fresh image was predicted as Not Fresh (Figure 3). In several samples, attribution lay around the eye and extended to the head, so the information used by the model cannot be restricted to the pupil. One correctly predicted Fresh example appeared predominantly blue, but gradient values cannot be inferred from the appearance of the overlay alone. The original images and the raw heatmap matrices were not available as separate outputs.

The samples consisted of three Fresh and three Highly Fresh images, with no actual Not Fresh class. The visualization therefore does not yet explain the attribution pattern of all classes or the dominant error of Not Fresh predicted as Fresh. The proportion of five correct predictions out of six images is not an estimate of model accuracy because the sample is small and not stratified. The visual findings are used to guide further examination, not to prove that the model has understood the biological indicators of freshness (Arun et al., 2021).

A more informative use of Grad-CAM in future work would combine class-stratified sampling, including correct and incorrect predictions of every class, with quantitative checks of the explanations. Such checks include sanity tests that randomize model parameters and deletion or insertion tests that measure how the predicted probability changes when the most highly attributed regions are removed or revealed (Arun et al., 2021). Combined with manual annotation of the eye region, these procedures would allow the overlap between attribution and the eye to be measured rather than judged visually.

Position of the Findings and Limitations

The results are in line with the use of pretrained representations for the visual assessment of fish in (Akgül et al., 2023; Yasin et al., 2023). The focus of this study lies in comparing configurations of a single backbone, whereas Prasetyo et al. (2022) developed an architectural modification on a different dataset. The accuracy of 97.22% is not used to claim superiority over those studies because their inputs, numbers of classes, units of separation, and evaluation procedures differ. The contribution that can be defended is evidence from an internal comparison and the documentation of the relationship between overall metrics, per-class errors, and visual attribution.

The pattern across configurations is also consistent with general findings on transfer learning. The weak performance of the frozen backbone (S1) agrees with the observation that deep pretrained features become increasingly specific to the source task and transfer less well to dissimilar targets (Yosinski et al., 2014), and with reports that the benefit of ImageNet pretraining varies across fine-grained domains (Kornblith et al., 2019); close-up images of fish eyes, in which the relevant cues are corneal opacity and pupil shape, belong to such a domain.

In contrast, Kim et al. (2022) concluded from medical imaging studies that pretrained networks used as feature extractors are often an efficient choice; the present results suggest that, for fish-eye images, at least partial adaptation of the deeper blocks is needed. Among fish freshness studies, Prasetyo et al. (2022) reported an accuracy of 63.21% on the eight-species Freshness of the Fish Eyes dataset, whereas Akgül et al. (2023) exceeded 96% on a binary task with two species and Yasin et al. (2023) reached 100% on a binary fish-body dataset.

The higher accuracies obtained here are more plausibly explained by the smaller number of species, the controlled acquisition conditions of MFED, and the likely presence of the same individuals in training and test partitions than by any superiority of EfficientNetV2B0. Finally, the concentration of S3 errors in Not Fresh images predicted as Fresh is consistent with the gradual progression of corneal clouding and pupil changes assessed in sensory schemes such as the Quality Index Method (Franceschelli et al., 2021).

The retrospective comparison of minimum val_loss also placed S3 in the lowest position. This agreement supports S3 as a candidate but does not remove the use of test results in the actual decision on configuration selection. Independent test data that have not influenced any decision are needed for confirmation. An audit of fish identity is also needed before linking the results to the ability to recognize new individual fish (Kapoor & Narayanan, 2023). From a practical perspective, the results suggest that a compact backbone such as EfficientNetV2B0 with partial fine-tuning may be a reasonable starting point for eye-based freshness screening tools, including applications intended for mobile devices.

Such a tool would still require controlled image acquisition, because the MFED images were captured under specific lighting, backgrounds, and angles, and its outputs would need to be framed as an estimate of the storage-based label rather than as a

guarantee of food safety. Field images taken in markets, with different species, lighting, and handling, are likely to differ from the training distribution and would need separate evaluation. Other limitations include a single seed, the absence of external testing, simultaneous changes in several training factors, and incomplete documentation of dependencies.

The computation time was not recorded, so efficiency across configurations can be compared only through the reconstructed numbers of trainable parameters. Repetition with several seeds and separation of training factors are needed to strengthen inference. Image manifests, environment versions, weights, and per-image predictions should be provided to support reproducibility and traceability of the experimental reporting (Heil et al., 2021; Pineau et al., 2021).

Table 8 summarizes which elements of the experiment were documented and which were not, following the reporting concerns raised in recommendations for improving reproducibility in machine-learning research (Pineau et al., 2021). The undocumented elements correspond directly to the limitations discussed above and indicate what should be added to the experimental records in a follow-up study.

Table 8. Documentation status of the experimental elements

Element	Status in this experiment
Dataset source, version, and licence	Documented (MFED version 2, CC BY 4.0)
Partition sizes and test class distribution	Documented; training and validation class distributions not printed; combined training and validation partitions contain 1,360 images per class (derived)
Partition procedure, duplicate audit, fish identity separation	Not documented; approximately 20 individuals inferred from the dataset description; fish-level re-split not yet performed
Preprocessing, augmentation, architecture, and head	Documented
Optimizer, learning rates, callbacks, and seed	Documented; one seed only
Hardware and software versions	T4 GPU metadata only; full dependency versions not recorded
Training duration and trainable parameters	Not recorded (duration); trainable parameters reconstructed from the architecture
Checkpoint selection and configuration selection	Checkpoints by val_loss; configuration for Grad-CAM by test macro F1; a retrospective validation-only rule selects the same configuration (S3)

Per-image predictions and raw heatmaps	Not stored; exact McNemar lower bound ($p \geq 0.125$) derived from aggregate counts

Source: Data Processed

CONCLUSION

This study compared four adaptation configurations of EfficientNetV2B0 for eye-based fish freshness classification on MFED version 2. Partial fine-tuning clearly outperformed both the frozen and the fully fine-tuned backbones. S3, which unfroze approximately the last 50% of layers excluding Batch Normalization, achieved 97.22% test accuracy and a macro F1 of 97.23% and was also selected by validation loss, but its advantage over S2 (96.67%) was not statistically significant. Validation and test rankings were identical.

Grad-CAM attributed predictions to the eye and surrounding head region, although Not Fresh images were not examined. In practice, S2 offers comparable accuracy with fewer than half of the trainable parameters, which favours lightweight screening tools. Because one seed was used and fish identity was not separated, confirmation on identity-separated data with several seeds is required.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study used a publicly available dataset (Multistage Fish Eyes Dataset version 2, licensed under CC BY 4.0) and a cloud-based computing environment (Google Colab), and no external financial support was received for the research, authorship, or publication of this article.

CONFLICT OF INTEREST

The author declares no conflict of interest. The author has no financial, commercial, institutional, or personal relationship with the creator of the Multistage Fish Eyes Dataset, the developers of the software and pretrained models used in this study (TensorFlow/Keras and EfficientNetV2B0), or any other party that could have inappropriately influenced the design, analysis, interpretation, or reporting of this research.

AUTHOR CONTRIBUTION

Tio Bagus Purnama Apriliano: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, and Writing – review & editing. As the sole author, Tio Bagus Purnama Apriliano designed the four adaptation configurations of EfficientNetV2B0, implemented the training, evaluation, and Grad-CAM pipeline, conducted the experiments and statistical analysis, prepared the tables and figures, and wrote and revised the manuscript. The author has read and approved the final version of the manuscript.

GENERATIVE ARTIFICIAL INTELLIGENCE USAGE STATEMENT

The author declares that generative artificial intelligence (AI) tools were used during the preparation of this manuscript to improve language clarity, readability, and grammatical structure, and, during revision, to assist in formatting citations and identifying candidate references. All references suggested with AI assistance were checked by the author against the original publications before inclusion. Generative AI was not used to design the study, generate or modify the data, produce the experimental results, figures, or tables, or draw the scientific conclusions. All AI-assisted text was reviewed and edited by the author, and its use complies with applicable ethical guidelines on the use and disclosure of generative AI in scientific publishing. The author takes full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

Akgül, İ., Kaya, V., & Zencir Tanır, Ö. (2023). A novel hybrid system for automatic detection of fish quality from eye and gill color characteristics using transfer learning technique. *PLOS ONE, 18*(4), e0284804. https://doi.org/10.1371/journal.pone.0284804

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., & Kalpathy-Cramer, J. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence, 3*(6), e200267. https://doi.org/10.1148/ryai.2021200267

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

FAO. (2020). *The State of World Fisheries and Aquaculture 2020: Sustainability in action*. FAO. https://doi.org/10.4060/ca9229en

FAO. (2024). *The State of World Fisheries and Aquaculture 2024: Blue Transformation in action*. FAO. https://doi.org/10.4060/cd0683en

Franceschelli, L., Berardinelli, A., Dabbou, S., Ragni, L., & Tartagni, M. (2021). Sensing technology for fish freshness and safety: A review. *Sensors, 21*(4), 1373. https://doi.org/10.3390/s21041373

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods, 18*(10), 1132–1135. https://doi.org/10.1038/s41592-021-01256-7

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports, 12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging, 22*(1), 69. https://doi.org/10.1186/s12880-022-00793-7

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2661–2671). IEEE. https://doi.org/10.1109/CVPR.2019.00277

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mohammadi Lalabadi, H., Sadeghi, M., & Mireei, S. A. (2020). Fish freshness categorization from eyes and gills color features using multi-class artificial neural network and support vector machines. *Aquacultural Engineering, 90*, 102076. https://doi.org/10.1016/j.aquaeng.2020.102076

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array, 16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research, 22*(164), 1–20. https://jmlr.org/papers/v22/20-303.html

Prasetyo, E., Purbaningtyas, R., Adityo, R. D., Suciati, N., & Fatichah, C. (2022). Combining MobileNetV1 and depthwise separable convolution bottleneck with expansion for classifying the freshness of fish eyes. *Information Processing in Agriculture, 9*(4), 485–496. https://doi.org/10.1016/j.inpa.2022.01.002

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 10096–10106). PMLR. https://proceedings.mlr.press/v139/tan21a.html

Yasin, E. T., Ozkan, I. A., & Koklu, M. (2023). Detection of fish freshness using artificial intelligence methods. *European Food Research and Technology, 249*(8), 1979–1990. https://doi.org/10.1007/s00217-023-04271-4

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems, 27*, 3320–3328.