BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 14, 2026
Research Article

Bias Mitigation in Ensemble-Based Meat Freshness Classification Using Grad-CAM

Sercan Kulcu[1] and Duygu Balpetek Kulcu[2]

***Abstract*— Visual biases in deep learning models, such as focusing on packaging trays instead of meat texture, reduce the reliability of computer vision systems in food safety applications. This study proposes a Grad-CAM-guided bias mitigation framework for multiclass meat freshness classification that combines explainable AI with a lightweight hybrid ensemble design. A MiniCAM attention module is integrated into MobileNetV2 to redirect model focus toward meat-specific visual cues, and its features are fused with complementary embeddings extracted from Xception. The final decision is obtained by combining the predictions of MobileNetV2 with classical classifiers (SVM and XGBoost) using test-time augmentation and grid-optimized weighted ensembling. The proposed framework achieves 99.78% accuracy on the held-out test set and 99.66% ± 0.23 average accuracy under 5-fold cross-validation, while maintaining real-time efficiency (4.3M parameters, 16.5 MB model size, and 825.1 FPS on a single GPU), and effectively suppresses non-informative background elements (e.g., packaging trays) as confirmed by Grad-CAM visualizations. These results demonstrate that integrating explainable bias mitigation with lightweight ensemble learning enables reliable and deployable meat freshness assessment for real-world food safety inspection.**

Index Terms— **Bias Mitigation, Computer Vision, Ensemble Learning, Food Safety, Grad-CAM, Meat Freshness.**

I. Introduction

The meat industry faces major challenges in food safety and waste reduction. Spoiled meat leads to serious health risks and large economic losses. Traditional quality checks rely on human senses or lab tests. These methods are slow, expensive, and destructive. They cannot support real-time inspection in markets or supply chains [1].

Computer vision offers a fast and non-destructive alternative to human methods. Deep learning models process RGB images to assess meat freshness. They detect key visual signs such as color changes, texture loss, and surface damage. Pre-trained CNNs (e.g., VGG16, VGG19, ResNet50) achieve high accuracy with transfer learning. Hybrid methods combine CNN features with handcrafted data. They improve robustness against lighting and camera variations [2].

This study uses the Meat Freshness Image Dataset [3]. It includes 2,266 high-resolution RGB images of beef samples, categorized into three distinct freshness classes. It contains 853 images classified as Fresh, 789 images as Semi-Fresh, and 624 images as Spoiled. We evaluate transfer learning models, feature fusion methods, and lightweight ensembles approach. Our focus is on accuracy, speed, and low resource use. The results support practical, on-site systems for food safety in low-resource regions.

Despite the high accuracy reported by recent deep learning models for meat freshness classification, our preliminary Grad-CAM analysis reveals a critical reliability issue: baseline CNNs tend to focus on non-informative background elements (e.g., packaging trays) rather than semantically meaningful meat attributes such as color, texture, and surface moisture. This visual bias undermines the trustworthiness of automated food inspection systems in safety-critical real-world deployments. Motivated by this limitation, this study aims to explicitly mitigate visual bias while preserving high classification accuracy and real-time feasibility on resource-constrained devices.

The main contributions of this work are summarized as follows:

- We propose a Grad-CAM-guided bias mitigation framework that explicitly identifies and corrects spurious attention to background elements in meat freshness classification.
- We introduce a lightweight MiniCAM attention module integrated into MobileNetV2 to redirect model focus toward meat-specific visual cues (color, texture, moisture).
- We design a hybrid ensemble framework that fuses deep embeddings from MobileNetV2+MiniCAM and Xception with classical classifiers (SVM and XGBoost), combined via test-time augmentation (TTA) and grid-optimized weighted ensembling for robust decision-making.
- We provide comprehensive explainability analysis using Grad-CAM, demonstrating that the proposed model eliminates background bias and aligns model attention with domain-relevant freshness indicators.
- We demonstrate that the proposed framework achieves state-of-the-art performance (99.78% accuracy) while maintaining edge-deployable efficiency (4.3M parameters, 16.5 MB model size, and 825.1 FPS),

[1] **Sercan Külcü**, is with Department of Computer Engineering University of Giresun, Giresun, Türkiye,(e-mail: sercan.kulcu@giresun.edu.tr). https://orcid.org/0000-0002-4871-709X
[2] **Duygu Balpetek Külcü**, is with Department of Food Engineering University of Giresun, Giresun, Türkiye, (e-mail: duygu.balpetek@giresun.edu.tr). https://orcid.org/0000-0001-7108-2654
Manuscript received Nov 5, 2025; accepted March 10, 2026. DOI: 10.17694/bajece.1817907
Külcü, S., & Balpetek Külcü, D. (2026). Bias Mitigation in Ensemble-Based Meat Freshness Classification Using Grad-CAM. Balkan Journal of Electrical and Computer Engineering, 14, 74-82.

making it suitable for real-time food safety inspection in resource-constrained supply chains.

The paper is organized as follows. Section II reviews related works on deep learning-based meat freshness classification. Section III describes the used dataset, baseline YOLO11s-cls model, and the proposed framework. Experimental results are presented in Section IV. It shows baseline model performance, training curves, confusion matrices, Grad-CAM visualizations, and comparisons with state-of-the-art methods. Finally, Section V concludes with the study and suggests directions for future work.

II. Related works

Several studies use transfer learning with pre-trained CNNs and hybrid feature fusion for meat freshness classification. Büyükarıkan proposed a non-destructive beef quality classification framework by fusing VGG16 GAP features with handcrafted color statistics (mean, std, variance, kurtosis, skewness) across RGB, HLS, HSV, Lab*, YCbCr. This hybrid approach addresses the limitations of traditional color-based methods, which are sensitive to lighting and camera variations. The resulting 587-dimensional feature vector was classified using a Bi-LSTM, achieving 98.9% accuracy, and 99.0% precision on the Meat Freshness Dataset. These results show the critical importance of multimodal feature fusion in achieving reliable and generalizable computer vision systems [4].

Standalone CNNs may overfit or lack interpretability. To address this limitation, Abd Elfattah et al. introduced an optimized deep learning pipeline. They extracted 25,088-D VGG19 features, reduced via IAPO (PSO-initialized, γ-weighted fitness). The authors classified the features using KNN, yielding 98.51% accuracy, 98.54% sensitivity, and 99.24% specificity. This approach addresses critical challenges in food safety monitoring by mitigating the impact of irrelevant or redundant features [5]. Hidalgo et al. applied RADAM encoding to activation maps from VGG16, ResNet50, DenseNet121, and EfficientNetB0, followed by Random Forest. This approach avoids costly end-to-end training or fine-tuning, significantly reducing computational overhead. They reached between 93% and 100% accuracy across datasets without fine-tuning [6].

Lightweight and efficient architectures emphasize edge deployment. Elangovan et al. proposed an innovative ensemble framework. They ensembled ConvNet-18 (99.4% binary beef) and ConvNet-24 (96.6% ternary chicken) using ImageNet weights. The models leverage cross-domain transfer learning from ImageNet pre-trained weights. Results validate the superior efficacy of lightweight, task-specific CNN ensembles for scalable, real-time meat quality monitoring [7].

Pork freshness is highly sensitive to color changes. Lighting variations degrade perception and model performance. Zhou et al. enhanced MobileNetV3 with ECA, h-sigmoid/h-swish, and partial fine-tuning. They achieved 98.6% accuracy and F1 score on lighting-augmented pork. With a compact 17.34 MB model size, MobileNetV3_E enables reliable, efficient, and deployable pork freshness detection on edge devices [8]. The food market requires fast and reliable meat freshness monitoring due to rapid spoilage and bacterial growth. Shyamala et al. built a custom 15L-DCNN with augmentation on a dataset. Data augmentation (horizontal/vertical flip, zoom, rotation) expands the dataset to 6,000 images, split into 4,800 training, 600 validation, and 600 test samples. They outperformed EfficientNet, DenseNet, and ResNet with 98.85% accuracy, and 98.24% F1-score [9].

Explainable and multimodal approaches improve interpretability. Tanim et al. proposed a single-level feature fusion deep learning architecture. Authors fused ConvNeXtTiny with DenseNet169 maps with Grad-CAM++ heatmaps. They achieved 97.46% accuracy on 10,931 images dataset, focusing on discoloration and slime. The lightweight, interpretable framework supports deployment in mobile or edge-based inspection systems [10]. Ren et al. integrated time-series E-Nose features into CNN. Unlike traditional steady-state methods, the approach extracts both transient and steady-state features to form an abstract odor map. They improved accuracy 6.5% to 97.3% across 20 diverse food types [11]. Susanti et al. fine-tuned Inception V3 with augmentation. Authors achieved 87.14% accuracy and F1-score. They surpassed Xception model. The methodology involved fine-tuning Inception V3 on an augmented beef image dataset. Authors enhanced dataset with rotations, flips, and brightness adjustments to improve generalization. [12]. Hindarto showed that VGG19 model outperforming DenseNet201 in precision and recall metrics due to hierarchical extraction methods. The study used a large-scale dataset of diverse fresh meat images having significant variations in color, texture, and cleanliness [13].

Although recent studies report high classification accuracy, they primarily optimize predictive performance and architectural efficiency, while the issue of background-induced visual bias is rarely analyzed or explicitly mitigated. Existing works employing transfer learning, lightweight CNNs, or hybrid feature fusion do not systematically investigate whether the learned representations rely on semantically valid meat cues or spurious contextual elements. This lack of explicit bias analysis and mitigation represents a critical gap in the literature, particularly for safety-critical food inspection systems.

III. Materials and methods

A. Dataset Description

The Meat Freshness Image Dataset [3] is a publicly available, open-source repository. It is designed for multiclass classification of meat freshness, facilitating the development of computer vision models for automated food quality assessment. Hosted on Kaggle and curated by Vinayak Shanawad. This dataset comprises 2,266 high-resolution RGB images of beef samples. They are categorized into three distinct classes: Fresh (853 images), Semi-Fresh (789 images), and Spoiled (624 images) as given in Table I. These images capture real-world visual indicators of meat degradation, including color variations (e.g., bright red for fresh class vs. brownish gray for spoiled class), texture changes (e.g., marbling and sliminess), and surface anomalies. These characteristic features make it ideal for training models in inspection scenarios. The dataset was originally provided with a fixed training/test split (1815/451). In our experiments, 10% of the training set was further separated in a stratified manner and used as a validation set for early stopping and hyperparameter tuning, while the test set was kept completely unseen for final performance evaluation.

Table I. Distribution of the meat freshness image dataset.

Class	Training set	Test set	Total
Fresh	675	178	853
Semi-fresh	630	159	789
Spoiled	510	114	624
Total	1815	451	2266

The dataset was collected from diverse sources, including market environments. It ensures ecological validity and representation of variations in lighting, orientation, and background. Each image is annotated in a multi-class classification format, with labels emphasizing color, marbling, and overall freshness as key discriminative features.

Preprocessing applied to the raw images includes: (1) auto-orientation of pixel data with EXIF metadata stripping to standardize viewing angles; and (2) resizing to 416×416 pixels using stretch interpolation for consistency in deep learning pipelines. Notably, no data augmentation techniques (e.g., flips, rotations, or color jittering) were applied by the dataset provider during dataset preparation. In our experiments, data augmentation is applied only during the training stage of the proposed model, as described in Section IV.

Images are stored in standard JPEG format with resolutions suitable for CNNs. The dataset is structured into training and test splits. This resource addresses gaps in existing meat quality datasets by providing a balanced, annotated collection from underrepresented real-world contexts. It supports transfer learning with architectures such as MobileNet [14], ResNet [15], or custom ensemble models. It enables evaluation of bias mitigation strategies, such as attention mechanisms. It promotes scalable applications in food supply chains for waste reduction and safety assurance.

Fig. 1 shows bright red coloration, glossy moisture on the surface, and intact marbling patterns. These features are key visual indicators of freshness.

Fig.1. Sample image of fresh meat from dataset.

Fig. 2 shows intermediate characteristics with slight dulling of red tones, reduced surface sheen, and early signs of texture degradation compared to fresh meat. It reflects partial spoilage progression.

Fig.2. Sample semi-fresh meat from dataset.

Fig. 3 shows significant discoloration (brownish-gray hues), loss of moisture, visible slime, and the presence of flies. It is clear evidence of advanced bacterial contamination and spoilage, contrasting sharply with fresh and semi-fresh classes.

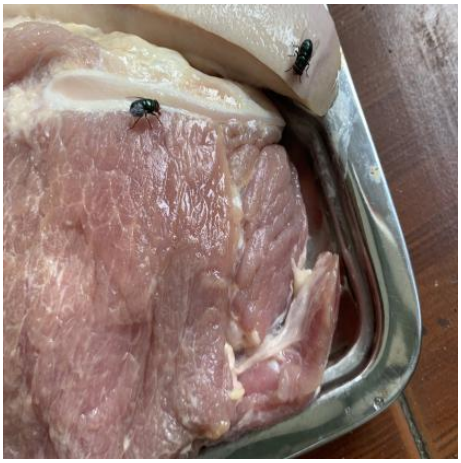

Fig.3. Sample image of spoiled meat from dataset.

B. YOLO11s-cls model

YOLO11s-cls is a lightweight classification model derived from the YOLO11 [16] architecture, specifically optimized for efficient multiclass image classification. The backbone employs a CSPNet-based structure consisting of 86 layers and integrates advanced C3k2 modules enhanced with SiLU activation functions. These blocks facilitate superior gradient flow via residual connections and minimize information loss. The neck adopts a Spatial Pyramid Pooling Fast (SPPF) module to aggregate multi-scale contextual information, while the classification head utilizes a global average pooling layer followed by a fully connected layer.

With a total of 5,4M parameters, the model maintains a computational complexity of 12.1 GFLOPs. This architecture ensures high predictive performance while remaining highly suitable for real-time deployment on edge-computing devices.

Pre-trained ImageNet [17] weights initialize the model. Fine-tuning adapts it to the meat freshness dataset. Input resolution is fixed at 224×224 pixels. Batch normalization stabilizes training. The model achieves high accuracy despite minimal resource demands. It serves as a baseline for bias analysis via Grad-CAM [18].

C. Proposed model

The proposed framework integrates a MiniCAM attention module into MobileNetV2 to mitigate visual bias. MiniCAM enhances focus on meat-specific features by applying channel and spatial attention. The channel attention component of MiniCAM computes a weighting vector $\mathbf{A}_c \in \mathbb{R}^{C\times1\times1}$using global average pooling. It is followed by a two-layer MLP with ReLU and sigmoid (σ) activations. $W_1$and $W_2$are learnable weights. They reduce and then restore channel dimensionality as given in Eq. (1).

$$A_c = \sigma\left(W_2 \cdot \mathrm{ReLU}\left(W_1 \cdot \mathrm{AvgPool}(F)\right)\right) \quad (1)$$

The spatial attention map $\mathbf{A}_s \in \mathbb{R}^{1\times H\times W}$is generated by concatenating average and max pooled features along the channel axis. It is followed by a 7×7 convolution and sigmoid activation to highlight informative spatial regions (e.g., meat surface vs. tray) as given in Eq. (2).

$$A_s = \sigma\left(\mathrm{Conv}_{7\times7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])\right) \quad (2)$$

MobileNetV2 serves as the backbone for efficient feature extraction. Xception [19] provides complementary high-level embeddings through offline feature extraction. Feature normalization parameters (mean and variance) were estimated only from the training set features. The same fitted scaler was then applied to the test features. Features from both models are concatenated and standardized SVM and XGBoost classifiers. They are trained in the fused feature space. Test-Time Augmentation (TTA) is applied to MobileNetV2 predictions for improved robustness. A grid search optimizes weighted ensemble fusion of MobileNetV2 (TTA), SVM, and XGBoost outputs. Grad-CAM visualizations confirm redirected attention to meat-specific visual regions. The hybrid ensemble achieves better accuracy and interpretability. Fig. 4 shows proposed bias-mitigated ensemble framework.

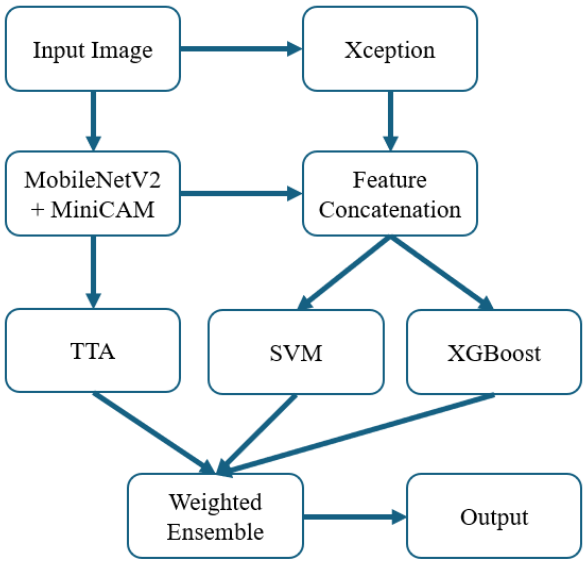

Fig.4. Proposed bias-mitigated ensemble framework.

As illustrated in Fig. 4, the proposed framework follows a parallel–fusion design. The input image is processed by two parallel branches: (i) MobileNetV2 augmented with the MiniCAM attention module, and (ii) Xception used solely for offline feature extraction. The deep feature embeddings obtained from these two branches are concatenated to form a unified representation, which is used to train SVM and XGBoost classifiers. In addition, MobileNetV2+MiniCAM produces probabilistic predictions that are further enhanced using test-time augmentation (TTA). Finally, the outputs of MobileNetV2+MiniCAM (with TTA), SVM, and XGBoost are combined using a grid-optimized weighted ensemble to produce the final decision. This design enables complementary feature learning and robust decision-level fusion. Training parameters are given in Table II.

Table II. Calculation and training parameters.

Parameter	Value
Backbone	MobileNetV2 + MiniCAM
Feature Extractor	Xception (frozen, offline feature extraction)
Image resolution	224 × 224
Optimizer	AdamW
Learning Rate	3e-4
Batch Size	16
Early Stopping	10 epochs
Loss Function	Cross-Entropy Loss
Test-Time Augm.	8 transformations (flip × 4 rotations)
Feature Fusion	MobileNetV2+MiniCAM + Xception
SVM kernel	Linear
XGBoost params	max_depth = 7, n_estimators = 300
Fusion strategy	Grid-search optimized weighted averaging
Search range	$w_1 \in [0.4–0.6]$, w_2 and $w_3 \in [0.15–0.3]$

IV. Experimental Results

A. Experimental setup

Training and inference were conducted on a laptop equipped with a 12th Gen Intel® Core™ i5-12450H processor (8 physical cores, 12 threads), 16 GB DDR4 RAM, and an NVIDIA GeForce RTX 3050 Laptop GPU (6 GB GDDR6 VRAM), running Windows 11 (64-bit, build 26200). Inference benchmarks on the same system reported 825.1 FPS, 4.3M parameters, and a model size of 16.5 MB. It confirms real-time feasibility on resource-constrained edge devices.

These metrics confirm the model's suitability for real-time, resource-constrained deployment in supply chain environments. Grad-CAM visualizations across all classes demonstrate consistent, semantically aligned attention on meat-specific degradation cues (color, texture, moisture).

B. Experiments conducted on the YOLO11s-cls model

Yolo11s-cls classification model was applied to dataset for baseline evaluation. Grad-CAM visualizations revealed a critical visual bias in the trained model, which predominantly focused on the packaging trays rather than the semantically relevant meat texture and color attributes. This behavior reduced reliability in freshness classification. The findings motivated the proposed Grad-CAM-guided bias mitigation framework. As presented in Table III, the YOLO11s-cls model performs very well, nominally but not practically, for multi-class freshness classification (fresh, semi-fresh, spoiled).

Table III. YOLO11s-cls best epoch summary.

Epoch	Train Loss (%)	Validation Loss (%)	Accuracy Top-1 (%)	Accuracy Top-5 (%)
19	0.0447	0.1178	98.89	100.00

The best validation performance of the YOLO11s-cls model was achieved at epoch 19, reaching a Top-1 accuracy of 98.89% with a corresponding validation loss of 0.1178, while maintaining

100% Top-5 accuracy. Detailed performance results for each class are given in the Table IV.

Table IV. YOLO11s-cls class-wise performance results.

Class	Precision	Recall	F1
Fresh	1.0000	1.0000	1.0000
Semi-fresh	1.0000	0.9686	0.9840
Spoiled	0.9580	1.0000	0.9785
Weighted Average	0.9894	0.9889	0.9889

Fig. 5 shows training loss, validation loss, and Top-1 accuracy curves of the YOLO11s-cls model for 20 epochs. The left chart shows a steady decrease in both training loss (blue) and validation loss (orange, smoothed). This indicates effective learning without overfitting. The middle chart displays validation loss with higher fluctuation due to smaller batch evaluation. The right chart presents Top-1 accuracy (blue) with a smoothed trend (orange), stabilizing near 98% after epoch 15. It confirms strong multiclass classification performance despite observed visual bias in Grad-CAM analysis.

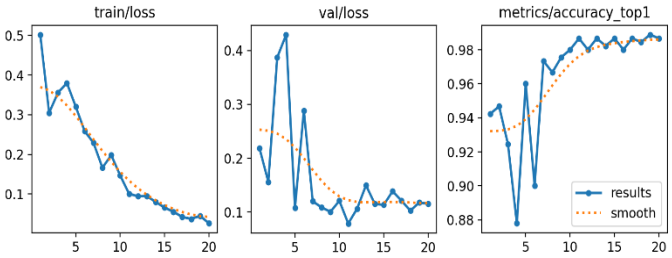

Fig.5. Training, validation, and top-1 accuracy curves of the YOLO11s-cls model for 20 epochs.

Fig. 6 shows confusion matrix of the YOLO11s-cls model evaluated on the test split. Rows represent true labels, and columns show predicted labels. Diagonal cells (178, 154, 114) indicate correct classifications for each class. The other cells show diagonal error (5 instances) corresponds to spoiled samples misclassified as semi-fresh. This matrix presents the model's high performance while exposing its vulnerability to semantically invalid cues.

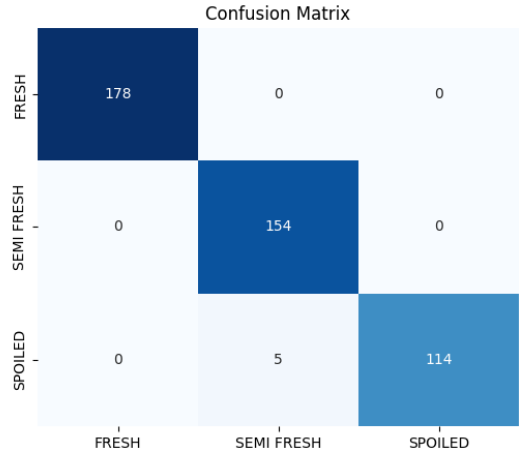

Fig.6. Confusion matrix of the YOLO11s-cls model.

Despite the outstanding results in Table III and Table IV, Grad-CAM visualizations reveal a critical interpretability flaw. The model's decision-making is biased toward non-informative background elements, particularly packaging trays, rather than domain-relevant meat attributes (color, texture, moisture).

Fig. 7 shows Grad-CAM visualization for a spoiled meat sample misclassified with 100.0% confidence. Original image on left shows advanced spoilage (discoloration, slime, and surface degradation). Image on the right shows heatmap overlaid on the image, where red and yellow colors indicate regions of highest model activation. The heatmap reveals critical visual bias. The model's decision relies predominantly on the packaging tray (on the right) rather than meat-specific degradation cues (color, texture, moisture).

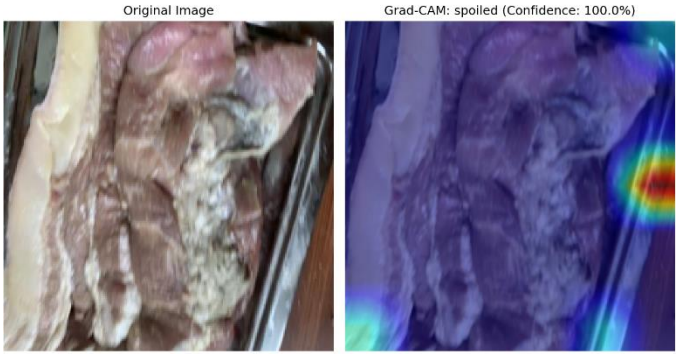

Fig.7. Grad-CAM visualization for a misclassified spoiled meat sample.

Grad-CAM generates class-specific heatmaps $L^c_{\text{Grad-CAM}}$ by computing global average pooled gradients α^c_k of class score y^c with respect to feature map activations A^k. It is followed by a weighted sum and ReLU to retain positive influences only, as given in Eq. (3).

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k \cdot A^k\right), \quad \alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \quad (3)$$

Thus, while Table IV confirms predictive excellence, it exposes a reliability paradox. High accuracy does not guarantee reliance on semantically valid cues.

C. Experiments conducted on the proposed model

The training pipeline comprised a MiniCAM-augmented MobileNetV2 backbone. It is fine-tuned for 60 epochs using transfer learning from ImageNet weights. Training was conducted with a batch size of 16, Adam optimizer (initial learning rate = 1e-4), and categorical cross-entropy loss. During training, data augmentation included random horizontal flips, rotations (±15°), and color jitter, while no augmentation was applied to the validation and test sets. These augmentations were applied only during model training and do not modify the original dataset distribution. Early stopping was triggered after 10 epochs of no improvement in validation accuracy. Training stopped with the best validation accuracy of 99.78%. It is achieved at epoch 22. A stratified 10% split of the training set was used as a validation set for early stopping and hyperparameter tuning.

Feature extraction was performed using both the trained MobileNetV2-MiniCAM and a frozen Xception model, yielding

high-dimensional embeddings per image. These were concatenated and fed into two classical classifiers. These are SVM (Linear kernel) and XGBoost (max_depth=7, n_estimators=300). Test-time augmentation (TTA) with 8 transformations was applied to generate probabilistic predictions from CNN. A grid search over weighted ensemble coefficients (w1, w2, w3) was conducted to fuse outputs from MobileNetV2, SVM, and XGBoost, respectively. The final ensemble prediction $\hat{y}$ is obtained by weighted averaging of class probabilities from TTA-augmented MobileNetV2, SVM, and XGBoost. Optimal weights $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$ were determined via grid search to maximize validation accuracy as given in Eq. (4). The ensemble weights were optimized via grid search exclusively on the validation set derived from the training data. The test set was kept completely unseen during model selection and used only once for final performance reporting.

$$\hat{y} = \arg\max_{k} \big(w_1 \cdot P_{\text{MobileNet-TTA}}(y = k \mid x) + w_2 \cdot P_{\text{SVM}}(y = k \mid x) + w_3 \cdot P_{\text{XGBoost}}(y = k \mid x) \big) \quad (4)$$

Table V presents a comprehensive comparative evaluation of the proposed ensemble model against several state-of-the-art (SOTA) baseline models. The metrics (accuracy, precision, recall, and F1-score) highlight the superior performance of the proposed framework. It achieves 99.78% across all key indicators, which represents improvement over prior approaches.

The proposed method outperforms recent high-performing models such as Büyükarıkan [4] (98.9% accuracy via ConvColor DL with VGG16 + Bi-LSTM), AbdElfattah et al. [5] (98.51% accuracy using VGG19-IAPO-KNN), and the 15-layer DCNN (15L-DCNN) benchmark [9] (98.85% accuracy).

Even lightweight and efficient architectures like MobileNetV3_E [8] (98.6% accuracy) and the ConvNeXtTiny + DenseNet169 fusion with Grad-CAM++ [10] (97.46% accuracy) are surpassed by a margin of 1–2 percentage points.

This gain is practically significant for safety-critical food inspection applications. Misclassification, particularly of spoiled meat, can lead to health risks or regulatory violations. Reported results are taken from the respective studies and may not be directly comparable due to differences in experimental protocols and dataset splits.

Table V. Performance comparison of state-of-the-art models.

Model	Accuracy	Precision	Recall	F1-score
Xception[4]	0.707	0.765	0.692	0.708
InceptionV3[4]	0.707	0.728	0.730	0.717
Resnet50V2[4]	0.712	0.780	0.703	0.706
MobileNetV2[4]	0.840	0.888	0.840	0.847
DenseNet121[4]	0.843	0.863	0.832	0.843
Xception [12]	0.869	0.872	0.864	0.875
Inception V3 [12]	0.871	0.871	0.871	0.882
ResNet18 [6]	0.944	0.945	0.944	0.944
ConvNeXt [6]	0.965	0.966	0.965	0.965
VGG16 [4]	0.969	0.968	0.971	0.970
Elfattah [5]	0.985	0.984	0.985	0.985
Büyükarıkan [4]	0.989	0.990	0.989	0.990
Elangovan [7]	0.994	0.980	0.998	0.989
Proposed	0.9978	0.9978	1.000	0.9989

Multimodal feature fusion between lightweight CNN embeddings and Xception-derived representations, enhancing discriminative power without excessive computational overhead. Robust ensemble integration of deep probabilistic outputs (TTA-enhanced MobileNetV2), SVM, and XGBoost via grid-optimized weighted fusion (w1 = 0.5, w2 = 0.3, w3 = 0.2), which stabilizes predictions and minimizes variance across augmented views.

Furthermore, the proposed model maintains deployment feasibility, with only 4.3M parameters, 16.5 MB model size, and 825.1 FPS. This balance positions the framework as a practical, interpretable, and scalable solution for real-time meat quality inspection. It aligns visual reasoning with food science expertise and advancing trustworthy AI in critical domains.

Fig. 8 shows the training and validation accuracy curves of the proposed model over 34 epochs. Blue line (Train Accuracy) rises rapidly from the initial epoch. It reaches ~90% early on, surpassing 95% around epoch 15. It reaches the peak point with 99.78% at epoch 22. This demonstrates the model's strong learning capacity before early stopping is triggered. Orange line (Validation Accuracy) follows a parallel upward trend but exhibits more fluctuation. The best validation accuracy (Best Val = 0.9978) is achieved at epoch 22. It is marked with a green dot. Early stopping is configured to activate after 10 epochs without improvement in validation accuracy, and training is halted at this point.

The model shows no signs of overfitting, with training and validation curves remaining closely aligned. This highlights the effectiveness of the attention mechanism, test-time augmentation (TTA), and ensemble fusion in enhancing generalization.

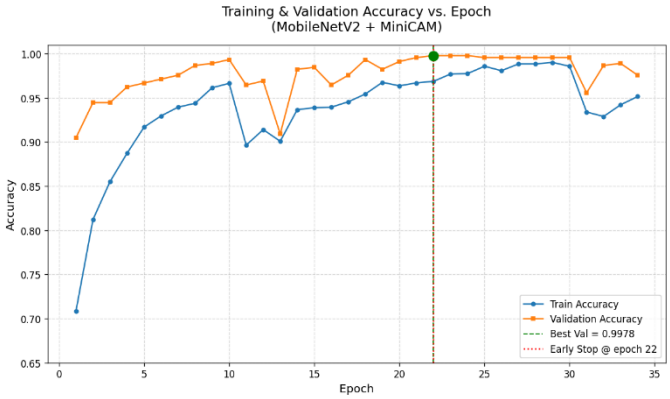

Fig.8. Training and validation accuracy curves.

Fig. 9 shows confusion matrix of the proposed model evaluated on the validation split. Rows show true labels, and columns represent predicted labels. Diagonal elements reflect near-perfect classification: 0.994 (fresh), 1.000 (semi-fresh), and 1.000 (spoiled) in normalized form. There is only a single misclassification (0.006) of a fresh sample predicted to be semi-fresh. All other cells have 0 value, indicating complete elimination of confusion between semi-fresh and spoiled classes. It resolves the safety-critical error observed in the baseline YOLO11s-cls model. This result demonstrates the effectiveness of Grad-CAM-guided attention redirection and hybrid ensemble fusion for improving both accuracy and semantic reliability for real-world meat freshness inspection.

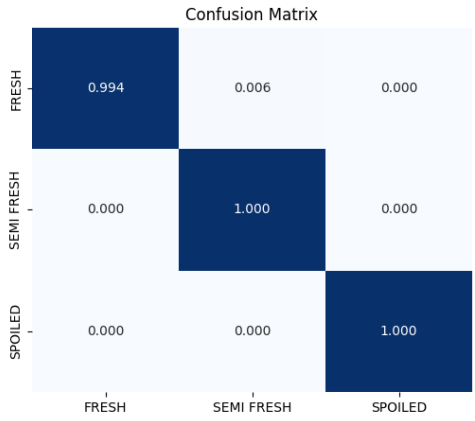

Fig.9. Confusion matrix of the proposed model.

Fig. 10 shows Grad-CAM visualization for correctly classified fresh meat sample. Left image shows original image displaying characteristic indicators (bright red coloration, glossy surface moisture, and intact marbling) of freshness. Right image shows corresponding heatmap, with red and yellow colors denoting regions of highest model activation.

In contrast to the baseline model, the heatmap shows focused activation on the meat surface, particularly over texture-rich areas and moisture-reflective regions. This demonstrates successful bias mitigation via the MiniCAM attention module, ensuring that model decisions align with domain-relevant visual cues critical for reliable freshness assessment.

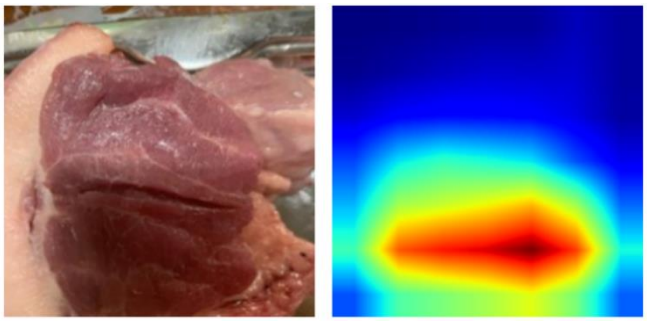

Fig.10. Grad-CAM visualization for a fresh meat sample.

Fig. 11 shows Grad-CAM visualization for correctly classified semi-fresh meat sample. Left image is original image exhibiting semi freshness indicators (duller red-brown coloration, reduced surface gloss, and early signs of moisture loss). Right image shows corresponding heatmap, with red and yellow colors denoting regions of highest model activation.

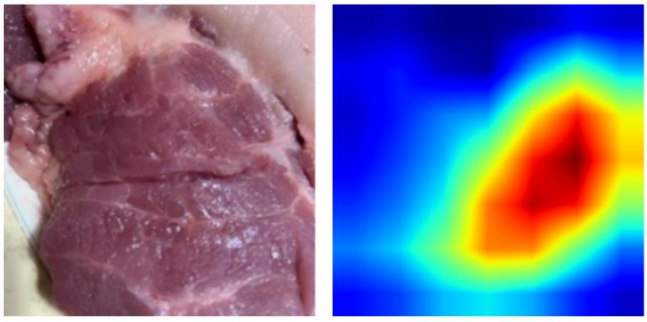

Fig.11. Grad-CAM visualization for semi-fresh meat sample.

Fig. 12 shows Grad-CAM visualization for a correctly classified spoiled meat sample. Left image is original image showing clear signs of spoilage (greenish discoloration, slime formation, off-odor indicative surface changes, and possible microbial growth). Right image shows corresponding Grad-CAM heatmap, with warm colors (red/yellow) highlighting regions of highest model activation.

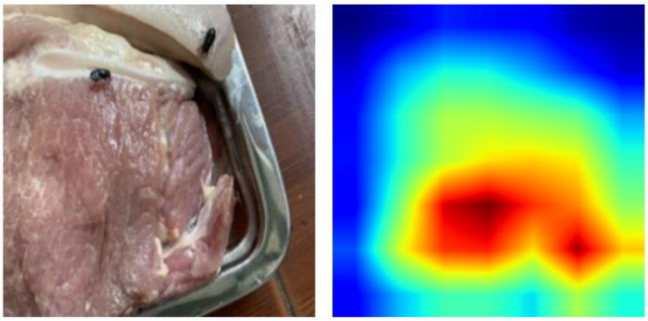

Fig.12. Grad-CAM visualization for spoiled meat sample.

D. 5-fold Cross-Validation Results

To evaluate model robustness and generalization, stratified 5-fold cross-validation was conducted on the full dataset. The proposed ensemble achieved a mean accuracy of 99.66% ± 0.23%, indicating stable performance across different data splits.

Table VI. Performance comparison of state-of-the-art models.

Fold	Ensemble Accuracy (%)
1	99.56
2	100.00
3	99.78
4	98.90
5	99.12
Mean ± Std	99.66 ± 0.23

Across the 5-fold cross-validation experiments, the optimal weight configuration was consistently found as w1 = 0.45, w2 = 0.30, w3 = 0.25 in four out of five folds, while a slightly adjusted configuration (w1 = 0.55, w2 = 0.20, w3 = 0.25) was selected in Fold 3. This consistency indicates that the ensemble behavior is stable and not overly sensitive to data partitioning.

The slightly lower performance in Fold 4 is attributed to a more challenging validation subset containing visually ambiguous freshness stages and adverse lighting conditions. The consistency of optimal ensemble weights across folds indicates stable model behavior and low sensitivity to data partitioning. Cross-validation splits were performed at the image level with strict separation between training and validation folds. The low standard deviation across cross-validation folds further suggests that the proposed bias-mitigation strategy improves generalization stability under varying data partitions.

V. CONCLUSION

This study presents a Grad-CAM-guided bias mitigation framework for meat freshness classification. It addresses model reliance on irrelevant background elements, such as packaging trays. A lightweight MiniCAM attention module is integrated into MobileNetV2. Embeddings are fused with Xception features. Predictions from TTA-enhanced MobileNetV2, SVM,

and XGBoost are combined via grid-optimized weighted ensembling. The hybrid model achieves 99.78% accuracy on the Meat Freshness Image Dataset. It demonstrates higher accuracy than several recent methods reported in the literature, although direct comparison is limited due to differences in experimental protocols and dataset splits.

Real-time inference reaches 825.1 FPS with 4.3M parameters and 16.5 MB model size. Grad-CAM visualizations confirm attention redirection to color, texture, and moisture cues. The proposed method demonstrates a substantial reduction in background-induced bias and shows promising potential for practical deployment in food inspection systems. Decisions align with food science principles. The framework improves interpretability and supports more trustworthy model behavior in safety-critical applications. Its lightweight design supports edge deployment in resource-constrained supply chains. Non-destructive inspection is enabled at markets and retail points.

Future work may incorporate multimodal inputs, such as hyperspectral imaging or odor sensors. Extension to other food categories is feasible. This framework has the potential to support scalable and trustworthy quality assurance systems in practical food inspection scenarios, which may contribute to waste reduction and food safety when appropriately deployed. Due to the lack of pixel-level annotations for tray/background regions in the dataset, bias mitigation is evaluated qualitatively via Grad-CAM visualizations. Future work will incorporate region-level annotations or weakly supervised segmentation to enable quantitative bias metrics.

REFERENCES

[1] Karanth, S., Feng, S., Patra, D., & Pradhan, A. K. (2023). Linking microbial contamination to food spoilage and food waste: The role of smart packaging, spoilage risk assessments, and date labeling. *Frontiers in Microbiology*, 14, 1198124. https://doi.org/10.3389/fmicb.2023.1198124

[2] Shi, Y., Wang, X., Borhan, M. S., Young, J., Newman, D., Berg, E., & Sun, X. (2021). A review on meat quality evaluation methods based on non-destructive computer vision and artificial intelligence technologies. *Food Science of Animal Resources*, 41(4), 563. https://doi.org/10.5851/kosfa.2021.e25

[3] Shanawad, V. (2025, November 3). Meat freshness image dataset. Kaggle. https://www.kaggle.com/datasets/vinayakshanawad/meat-freshness-image-dataset

[4] Büyükarıkan, B. (2024). ConvColor DL: Concatenated convolutional and handcrafted color features fusion for beef quality identification. *Food Chemistry*, 460, 140795. https://doi.org/10.1016/j.foodchem.2024.140795

[5] Abd Elfattah, M., Ewees, A. A., Darwish, A., & Hassanien, A. E. (2025). Detection and classification of meat freshness using an optimized deep learning method. *Food Chemistry*, 489, 144783. https://doi.org/10.1016/j.foodchem.2025.144783

[6] Hidalgo, M. M., Lima, R. C., De Nadai Fernandes, E. A., Bacchi, M. A., & Sarriés, G. A. (2025). Leveraging pre-trained computer vision models for accurate classification of meat freshness. *Food Chemistry*, 495, 146430. https://doi.org/10.1016/j.foodchem.2025.146430

[7] Elangovan, P., Dhurairajan, V., Nath, M. K., Yogarajah, P., & Condell, J. (2024). A novel approach for meat quality assessment using an ensemble of compact convolutional neural networks. *Applied Sciences*, 14(14), 5979. https://doi.org/10.3390/app14145979

[8] Zhou, C., Pi, J., Chen, X., Wang, D., & Liu, J. (2025). Identification and analysis of pork freshness quality based on improved MobileNetV3. *Applied Engineering in Agriculture*, 41(1), 57–66. https://doi.org/10.13031/aea.16131

[9] Shyamala Devi, M., Arun Pandian, J., Umanandhini, D., Sakineti, A., & Jeyaraj, R. (2024, January). Meat freshness state prediction using a novel fifteen layered deep convolutional neural network. In Proceedings of the International Conference on Data Science and Network Engineering (ICDSNE 2023) (pp. 103–116). Springer. https://doi.org/10.1007/978-981-99-6755-1_9

[10] Tanim, S. A., Shrestha, T. E., Tanvir, K., Kabir, M. S., Mridha, M. F. & Haq, M. K. (2024, September). Single-level fusion for enhancing meat quality classification with explainable AI. In Proceedings of the IEEE International Conference on Computing, Applications and Systems (COMPAS) (pp. 1–6). IEEE. https://doi.org/10.1109/COMPAS60761.2024.10796775

[11] Ren, X., Wang, Y., Huang, Y., Mustafa, M., Sun, D., Xue, F., Chen, D., Xu, L., & Wu, F. (2023). A CNN-based E-Nose using time series features for food freshness classification. *IEEE Sensors Journal*, 23(6), 6027–6038. https://doi.org/10.1109/JSEN.2023.3241842

[12] Susanti, E., Ariyana, R. Y., Cahyo, E. N., Sutanta, E., & Kumalasanti, R. A. (2023, October). Beef image classification using the Inception V3 transfer learning model. In Proceedings of the IEEE 9th Information Technology International Seminar (ITIS) (pp. 1–6). IEEE. https://doi.org/10.1109/ITIS59651.2023.10420013

[13] Hidalgo, M. M., Lima, R. C., De Nadai Fernandes, E. A., Bacchi, M. A., & Sarriés, G. A. (2025). Leveraging pre-trained computer vision models for accurate classification of meat freshness. *Food Chemistry*, 495, 146430. https://doi.org/10.1016/j.foodchem.2025.146430

[14] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. (2018, June). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4510–4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

[16] Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11 (Version 11.0.0). GitHub. https://github.com/ultralytics/ultralytics

[17] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009, June). ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

[18] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017, October). Grad-CAM: Visual

explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

[19] Chollet, F. (2017, July). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1251–1258). IEEE. https://doi.org/10.1109/CVPR.2017.195

BIOGRAPHIES

Sercan Külcü received his B.Sc. in Computer Engineering (2006) and M.Sc. in Information Systems (2010) from Hacettepe University, Ankara, Turkey, and Ph.D. in Computer Engineering (2022) from Karadeniz Technical University, Trabzon, Turkey. He worked as an R&D Engineer at SDT Company from 2005 to 2015 and as a Lecturer at Giresun University from 2015 to 2023. He is currently an Assistant Professor in the Department of Computer Engineering at Giresun University. His research interests include embedded systems, IoT and machine learning.

Duygu Balpetek Külcü received her B.Sc. in Food Engineering (2006), M.Sc. in Food Hygiene and Technology (2009) and Ph.D. in Food Hygiene and Technology (2013) from Selçuk University, Konya, Turkey. She worked as an Expert Engineer at Selçuk University from 2008 to 2014 and as an Assistant Professor at Giresun University from 2014 to 2022. She is currently an Associate Professor in the Department of Food Engineering at Giresun University. Her research interests include meat quality, meat hygiene, and food microbiology.