Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

DASS-Net: A Multi-Branch Aesthetic–Spectral–Semantic Deep Learning Model for Food Spoilage Detection in Hospitality Operations

Randy O. Descarten[1], Jasten Keneth D. Treceñe[2]

[1]Hospitality Management Department, College of Business and Technology, Surigao del Norte State University, Surigao City, Philippines.
[2]Information Technology Department, College of Engineering, Eastern Visayas State University, Tanauan, Philippines.

E-mail: [1]randydescarten@gmail.com, [2]jastenkenneth.trecene@evsu.edu.ph

Abstract

Food freshness and food safety in the hospitality industry are, to date, mainly dependent on visual observation, where human visual interpretation does not always successfully identify the less obvious signs of early food spoilage. In this work, we introduce DASS-Net, a novel hybrid deep network with a trilinear structure and three separate task-focused branches: (1) the RGB aesthetic branch, (2) the LAB spectral branch, and (3) the Vision Transformer ViT-Small semantic branch, used together for simultaneous analysis of food image-based texture degradation, color deviation, and overall food image-related semantic dependencies on a global scale. Each of the network's task-focused branches projects its outputs to a unitary latent space and further refines them with multi-head attention and adaptive joint processing with a learnable gated fusion module. To train DASS-Net, a dual-objective training approach with class-weighted and smoothed Cross-Entropy loss and Contrastive Learning with MixUp and CutMix data augmentations is used. Eight-class classification experiments on fresh and spoiled food samples of bread, dairy products, fruits, and vegetables are conducted to evaluate the performance of DASS-Net, providing a validation accuracy of 94.33%, which demonstrates a 23.18% and 25.83% relative improvement over ResNet-18 with LAB and RGB visual channels, respectively. Additionally, the classification model yielded a minimum misclassification rate of 13.30% for spoiled items identified as fresh, which clearly marks an important characteristic related to safety issues. The complementarity between the spectral and aesthetic features has also been proven by the Grad-CAM visualizations of the dual-branch network, where the most confusion occurs in the case of dairy products with negligible surface decay, as revealed by the analysis of failure cases. The results explicitly confirm that the developed DASS-Net delivers a safe, efficient, and scalable vision solution applicable to kitchen intelligence and hospitality control scenarios.

Keywords: Food Spoilage Detection, Hospitality Management, Vision Transformer, Spectral–Aesthetic Fusion, Supervised Contrastive Learning.

DOI: https://doi.org/10.36548/jiip.2025.4.021
Received: 05.11.2025, received in revised form: 04.12.2025, accepted: 20.12.2025, published: 31.12.2025

1. Introduction

Presentation and quality of the food are considered two of the most critical aspects of the beliefs and expectations of hotels and restaurants since appearance can be taken as a vital indicator of food quality. Today, some hotel and restaurant guests judge food quality by sight, taste, or by reading its description. Regarding this, early spoilage identification has introduced a new factor in the hotel environment where food safety, presentation, and service excellence are combined in one place beyond functional requirements. This also safeguards consumer satisfaction, prevents food-borne diseases, and protects reputation [2]. Therefore, the ability to precisely and reliably assess the freshness of food is key to managing good hospitality service.

Traditional methods in hotels and restaurants include the use of human senses for spoilage detection in an extremely manual process for inferring whether foods are safe for human consumption. Food inspection for spoilage, color, texture, wetness, or damage to the surface of foods is conducted by chefs and hotel/restaurant kitchen staff. To a great extent, though, such methods have been very valuable, they are extremely prone to errors and depend on the personal skill level of the individual involved in conducting such processes [3]. This is especially important for restaurants where numerous ingredients are handled within a single day. In this way, such inspections would be inaccurate. In addition, low skill levels of staff, the amount of workload, and environmental factors could induce risks of minor initial signs being missed.

As more diversity emerges in food production in hotels, requiring more time sensitivity, it is even more apparent that there is a limitation in the process of merely observing it food through visual and manual means. Typically, this occurs in buffet meals, fast turnaround diets, and round-the-clock operations. Industry accounts indicate that developing, would-be spoilage, commonly referred to as those changing gradually on a continuum, often escapes detection during conventional hotel checks [4]. Since early warning signs of spoilage are not picked up, greater risks of lowering customer satisfaction, degrading food quality, accelerating hotel food waste, and even non-compliance with food safety standards readily emerge [5]. Consequently, there is an urgent need for creating a mechanism that can be scaled up, is objective, and automatic, hence assisting hotel personnel in making determinations regarding food freshness.

With the latest developments in computer vision, various machine learning and image-processing-based techniques have been demonstrated to be effective for certain food safety applications and automating food quality assessment. One of the techniques that showed very strong potential in the detection of spoilage across a wide range of food categories is the Convolutional Neural Network (CNN). CNNs learn complex spatial and colour-based patterns directly from raw images without requiring manual feature engineering. They also automatically learn hierarchical visual features, such as browning, textual degradation, and surface irregularities. These developments show that automated systems can exceed the consistency of manual inspections traditionally performed by kitchen staff.

Moreover, many previous studies from relevant areas, such as computer vision and food science, have examined various hand-crafted features like color histograms, signatures, or textures to represent visible changes associated with food spoilage [7]. Typically, they aim to examine each component of food deterioration in particular, such as dehydration processes, enzymatic browning, and bacterial growth. Although highly efficient in respective tasks, low-level spectral feature-based approaches overlook more general aesthetic image characteristics within food items, possibly including sharpness, color balance correction, believability of

depicted food composition on the plate, light reflections, and symmetry of depiction [8]. It appears that these characteristics are precisely what professional kitchen inspectors typically focus on during real-life tasks. A joint and rather tight focus on them provides a very low ability to detect the initial stages of food decomposition processes that attempt to impair the overall visual representation of food items, rather than merely discolored decay and mushrooms.

However, owing to this recent innovation in computer vision, several learning algorithms, along with various processes in images, have been identified as having the capacity to effectively and efficiently fulfill the duties of food safety and conduct food quality analysis. Among these identified processes that can assess food spoilage across all forms of food are Convolutional Neural Networks. These networks have been shown to have the ability to learn patterns across raw images, including browning and textural damage, which must be identified manually by personnel in the kitchen.

Models of aesthetic image analysis have developed methods capable of estimating the aesthetic quality of images based on composition, clarity, harmony, and stylistic properties [9]. Though such techniques are commonly employed in photography, social media content assessment, and automated visual scoring, they have not been comprehensively utilized in food quality and safety evaluation. In addition, a combination of aesthetic analysis and spectral analysis would greatly improve the sensitivity of detecting aesthetic properties, which decrease over time and indicate the early stages of food deterioration through reduced brightness, blurred edges, skewed color distribution, and reduced gloss.

The vast majority of existing work on food spoilage classification models involves single-stream architectures that discern either spectral information or textural characteristics separately. There has also been a lack of consideration for the benefits that can be gained from combining various visual aspects into one model. Existing work has not attempted to focus on the combination of aesthetic information and spectral characteristics involved in judging the freshness of foods, despite indications that both aspects can be perceived manually and automatically.

Moreover, most of these studies are more interested in applying these concepts in agricultural, industrial, or retail supply. However, relatively little research is currently being done on applying these concepts to food spoilage classification, wherein food safety and food quality are given equal importance.

1.1 Objectives of the Study

This research aims to propose and assess the effectiveness of a hybrid multi-branch deep learning approach that has the ability to perform accurate classification tasks of food spoilage from images within the hotel industry.

This research is keen on exploring the following aspects:

- Designing an aesthetic-spectral feature MAB with multiple branches, including the extraction of complementary features of an image within both RGB space and LAB.
- Adding a semantic Transformer branch to improve the perception of high-level patterns of spoilage.

- Applying the multi-head attention mechanism and gated fusion strategy for the optimal fusion of feature representations from the branches.
- Applying supervised contrastive learning to improve the separation between classes.
- Comparing the hybrid model with the model based on just RGB and with the model based on just LAB through an ablation study.
- Recommending a suitable model by taking into account accuracy, F1 scores, confusion matrices, and other performance indicators related to errors concerning spoiled-as-fresh products.
- Producing Grad-CAM attributions and studying failure cases to increase model interpretability.

1.2 Key Contributions

- The study presents a new hybrid model that combines RGB aesthetic features, LAB spectral cues, and high-level semantic representations using a Vision Transformer (ViT) branch. This design uses complementary information across color spaces and transformer-based context modeling. It creates a better feature extractor for detecting food spoilage.
- DASS-Net employs dynamic multi-head attention and a gated fusion mechanism to dynamically adjust the reliance on RGB, LAB, and transformer features for each input. This helps the model remain reliable even when lighting, textures, or food surfaces change, which often causes problems for single-stream CNNs.
- The model integrates Supervised Contrastive Learning (SupCon) to strengthen inter-class boundaries and reduce confusion between visually similar spoiled and fresh items. Moreover, the combined use of MixUp and CutMix across multi-branch inputs enhances regularization and improves generalization for limited datasets in hospitality settings.
- Dual-stream Grad-CAM visualization shows how the RGB and LAB branches focus on different signs of spoilage. The paper also introduces a safety-critical metric called the “spoiled-as-fresh rate” to measure high-risk misclassifications, which goes beyond accuracy and helps hotels and restaurants use the system safely.
- Enhances food safety and operational reliability to prevent both foodborne risks and unnecessary waste.
- Provides an automated, scalable, and deployment-ready vision-based system that supports high-volume hotel and restaurant workflows, offering continuous quality monitoring suitable for modern hospitality environments and smart kitchen platforms.

2. Related Work

Previous works on food spoilage detection show several methodological and practical limitations. These limitations pose challenges in practical applications in hospitality and kitchen environments. Most studies rely on single-stream CNNs using RGB images. However, RGB alone is insufficient since early spoilage involves subtle spectral shifts not visible in RGB, and different foods exhibit naturally varied colors that mask discolorations. Moreover, texture and moisture changes are not captured well by RGB-only models. Furthermore, there is limited use of alternative color spaces despite proven spectral advantages. Although few works have explored LAB and HSV features, LAB is rarely fused with deep learning architectures. Also, past color-space methods lacked robust model fusion, and spectral cues were used in isolation and did not combine with RGB or transformers.

Moreover, the majority of the studies have reported model accuracy only, such as validation, precision, recall, and F1-score; however, safety-critical errors, such as the spoiled-as-fresh rate, are not commonly considered. Correspondingly, real-world kitchens have varied lighting and backgrounds, whereas prior models often use clean datasets with controlled conditions, focusing only on fruit or milk-specific datasets, and have limited variation in illumination and noise. This reduces generalizability when applied to busy hotel kitchens. On the other hand, existing models treat samples independently and lack global reasoning. CNNs perform well at local features; however, they struggle with global relationships such as color gradients across the whole food surface. Similarly, they also struggle with structural changes and fine-grained decay progression.

No prior work integrates aesthetic cues (RGB), spectral cues (LAB), and high-level semantic representations (transformer-based models) into a unified architecture designed specifically for food safety applications. Additionally, advanced strategies such as MixUp, CutMix augmentation, and supervised contrastive learning are rarely used in food spoilage research despite their proven ability to improve robustness under varied lighting, texture, and surface conditions. Thus, the proposed Dual-Aesthetic-Spectral-Semantic Network (DASS-Net) uniquely integrates multi-modal visual cues, transformer-based semantic understanding, dynamic fusion mechanisms, and risk-centered evaluation to overcome the limitations of existing methods. Table 1 summarizes key studies on food spoilage and quality detection that outline their methods, performance, and limitations.

Table 1. Summary of Related Work on Food Spoilage Classification

Author(s)	Method	Performance	Limitation
[10]	CNN-based mould detection of bread microscopic images	F1-score of 0.9948	Exclusion of problematic images and datasets was taken on controlled imaging conditions. Basic data augmentation was applied.
[11]	Detection of mould on Food Surface using YOLOv5	Precision (98.10%), Recall (100%), Average Precision (99.60%)	Signs of overfitting. Trained only on heterogeneous images of surface-level mold and may not generalize to diverse food types, imaging conditions, and early-stage/low-visibility mold.

[12]	Near-infrared Hyperspectral Imaging in detecting spoilage in sausage	R^2 = 80.61%	Processed meat varies in color, and there are limited dataset types. The study shows that color-based spoilage detection is feasible, but limited.
[13]	Least-Squares Support Vector Machines (LSSVM), Competitive Adaptive Reweighted Sampling (CARS), and Interval Variable Iterative Shrinkage Approach (iVISSA) for Tan sheep using Hyperspectral Imaging	R^2 = 91.50%	Complex equipment and model. There is a need for a lightweight and camera-based approach that is more practical for kitchens and restaurants.
[14]	ResNet-50 for fruit decay detection	Validation Accuracy (98.89%)	Struggles with partially spoiled fruits and varying backgrounds
[15]	Naïve-Bayes Classifier for Milk Spoilage	Validation Accuracy (92.2%)	Dairy shows subtle texture changes. No safety-focused metrics.

3. Proposed Work

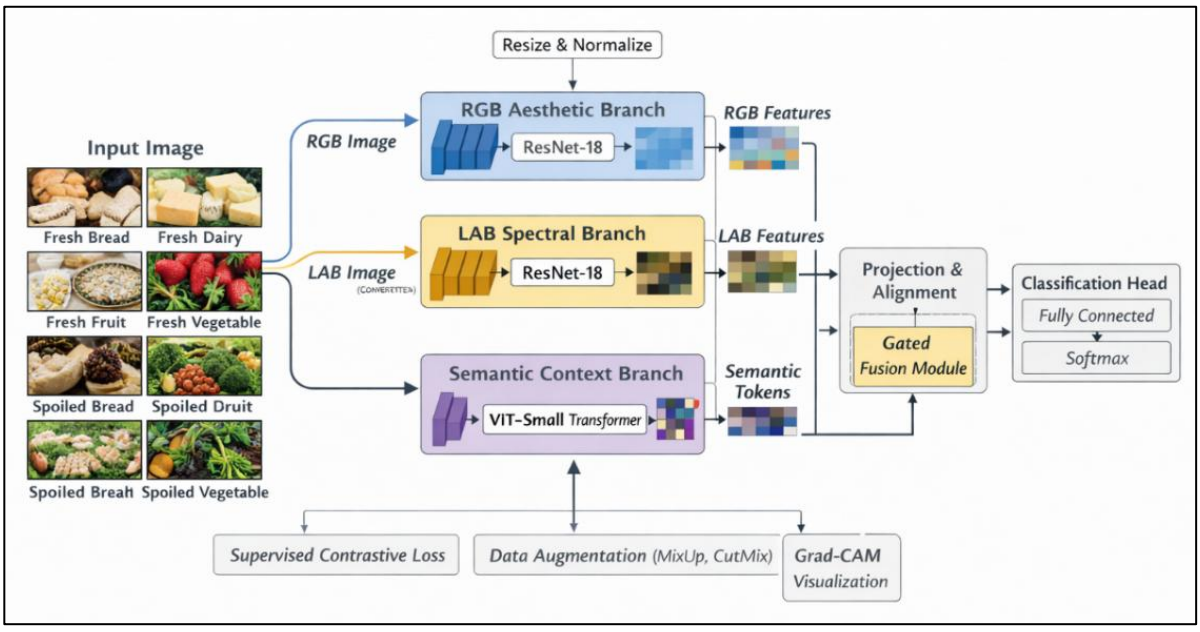


Figure 1. DASS-Net Architecture

As shown in Figure 1, the architecture of the DASS-Net used in the current study focuses mainly on the classification of food spoilage from images. The DASS-Net takes an RGB image as input, which then undergoes a resizing process; the normalized image is copied into the RGB stream and the LAB conversion of the spectral stream, with the original RGB

image proceeding to the semantic branches as well. The goal of the semantic branches of the ResNet-18 focuses on the texture/aesthetic surface, while the LAB branches also seek to find the spectral color discolorations/degradation. An additional goal of the Vision Transformer focuses on the global semantic context as well as the spoilage phenomena that are unclear to human visual observation. The three obtained features are then unified into a latent space; this space is improved by multi-head attention and adaptive fusion using the gated fusion module, with the outcome proceeding to the fully connected classifier head, which produces the final genre with the Softmax output. Grad-CAM visualizations provide explanations for the network's ability to perform balanced augmentation techniques and contrastive supervised learning during training.

3.1 Dataset and Pre-processing

3.1.1 Data Organization and Stratified Splitting

In this study, the eight marked classes of fresh and spoiled varieties are used as the dataset. The Fresh and Spoiled Food Image Dataset, from which the dataset for this study was acquired,can be found at the Kaggle repository [17]. In this dataset, the images of food from hotels in eight classes are in the RGB range,

$$C = \{fresh_{bread}, fresh_{dairy}, fresh_{fruits}, fresh_{vegetables}, spoiled_{bread}, spoiled_{dairy}, spoiled_{fruits}, spoiled_{vegetables}\}$$

Each class was kept in a separate directory, and the images inherited the label of the directory. Let N_c be the number of images belonging to class $c \in C$. Further, a stratified split per class was done using a fixed random seed so that the empirical class distribution was preserved across the training, validation, and test sets. For each class, 70% of the images went into the training set, 15% went to the validation set, and 15% into the test set. Next, the three subsets were independently shuffled in order to avoid any ordering bias. Before data splitting, the directory labels were manually inspected to minimize inconsistencies in the annotation. Again, stratified sampling with a fixed random seed preserved class distribution, while shuffled partitioning reduced the chances of temporal ordering leakage.

3.1.2 Image Pre-processing and Augmentation

All images went through a combined transformation step in a jointly learned augmentation graph equal for the RGB and LAB channels. The training transformation T_train was a combination of the following: (i) resizing the shorter side to a maximum of 1.1 × 224 pixels, (ii) applying a RandomResizeCrop of size 224 × 224 with a scale in [0.7, 1.0] and an aspect ratio in [0.9, 1.1], (iii) randomized horizontal flipping with respect to a 0.5 probability, (iv) randomized rotation in the range of [-20°, 20°], (v) Color Jittering for brightness, contrast, saturation, and hue, (vi) applying a Gaussian blur, (vii) auto-contrast, and (viii) applying random grayscaling. While training and evaluating the model, another transformation was used to resize the images to a fixed size of 224 × 224 to avoid any randomness in model testing and training. The fixed size was a trade-off between the representation and efficiency of the model. The model followed the standard ImageNet model practice on input size to maintain efficiency and representation balance. After augmentation, each image $I' \in [0.255]^{H \times W \times 3}$ was converted to a floating-point tensor $x^{rgb} \in \mathbb{R}^{3 \times 224 \times 224}$ and normalized using the standard ImageNet statistics, using Equation 1,

$$x_c^{rgb}(u,v) = \frac{\frac{I'_c(u,v)}{255} - \mu_c}{\sigma_c}, \tag{1}$$

$$\mu = (0.485, 0.456, 0.406),$$

$$\sigma = (0.229, 0.224, 0.225).$$

The Random Erasing operation (with a probability of 0.25) was introduced in the RGB tensor by overlaying rectangles with randomly selected pixels, which greatly assisted in enhancing robustness to occlusions and missing information. The thought process behind this design decision was that greater emphasis was placed on maintaining color integrity by employing the Random Erasing operation selectively on the input RGB images, which aimed to simulate a realistic occlusion scenario.

The values of ColorJitter parameters are kept within limited ranges to mimic practical lighting changes rather than simulate deterioration. Augmentation is performed symmetrically on both fresh and spoiled samples to avoid bias. While no negative impacts were observed in experiments, a detailed augmentation sensitivity analysis in the future would be useful for a proper understanding of the interaction between augmentations and semantics. No performance reduction was observed in the experimental evaluation, since orientation-based augmentation, such as horizontal flipping, was retained considering that spoilage visual cues are inherently orientation-independent [28].

3.1.3 RGB-LAB Color-Space Conversion

In parallel with the RGB branch, each augmented image was converted into the CIE-LAB color space to emphasize perceptually meaningful chromatic differences associated with food spoilage. The RGG array I' was first normalized to $[0, 1]$ and transformed to the CIEXYZ space using the standard linear transformation. The $L^{\backslash *}, a^{\backslash *}, b^{\backslash *}$ components were then computed as in Equation 2, Equation 3, and Equation 4,

$$L^{\backslash *} = 116 f\left(\frac{X}{Y_n}\right) - 16, \tag{2}$$

$$a^{\backslash *} = 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{X}{Y_n}\right)\right] \tag{3}$$

$$b^{\backslash *} = 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right] \tag{4}$$

where X_n, Y_n, Z_n denote the reference white point and $f(t)$ is the usual CIE piecewise function. For numerical stability and to bring all channels to a comparable range, LAB values were normalized as in Equation 5,

$$L' = \frac{L^{\backslash *}}{100}, a' = \frac{a^{\backslash *}}{128}, b' = \frac{b^{\backslash *}}{128} \tag{5}$$

producing a LAB tensor $x^{lab} \in \mathbb{R}^{3 \times 224 \times 224}$. This tensor was used as input to the spectral branch. RGB-LAB conversion followed standardized CIEXYZ procedures with fixed white-reference points and normalization scaling to comparable perceptual ranges. The

clipping safeguards have prevented the out-of-range anomalies to ensure that spectral variations have reflected natural spoilage chromatic shifts rather than computational artifacts.

3.2 DASS-Net Hybrid 3-Branch Architecture

3.2.1 Brach Encoders

The proposed DASS-Net combines a triple-cascade structure with a triplet loss objective for distinguishing between real and generated pairs. The system consists of a triple-cascade network with a triplet loss objective that combines three feature extraction branches; the RGB image branch, the LAB image branch, and the semantic transformer image branch. ResNet-18 was selected for the RGB image branch and LAB image branch due to the established stability demonstrate in the context of spectral learning with moderate computational costs that exhibit better robustness than lightweight architectures [29][30]. The chosen model thereby ensures that execution is possible even in kitchen environments.

3.2.1.1 Aesthetic RGB Branch (ResNet-18)

The RGB tensor x^{rgb} was fed to a ResNet-18 backbone f_{aes} configured for three-channel LAB input using Equation 6,

$$h_{aes} = f_{aes}(x^{rgb}) \in \mathbb{R}^{d_{aes}}, d_{aes} = 5 \tag{6}$$

3.2.1.2 Spectral LAB Branch (ResNet-18)

The LAB tensor x^{lab} was processed by a second ResNet-18 f_{spe} configured for three-channel LAB input using Equation 7,

$$h_{spe} = f_{spe}(x^{lab}) \in \mathbb{R}^{d_{spe}}, d_{spe} = 5 \tag{7}$$

This branch captured subtle color shifts related to browning or mold growth that may be less apparent in RGB.

3.2.1.3 Semantic Transformer Branch (ViT-Small)

To model higher-level semantic patterns, a Vision Transformer (ViT-Small) with a patch size of 16 f_{sem} processed the RGB tensor and represented it in Equation 8,

$$h_{sem} = f_{sem}(x^{rgb}) \in \mathbb{R}^{d_{sem}}, d_{sem} = 384. \tag{8}$$

Each branch feature was projected into a shared latent space of dimension D through separate linear layers followed by dropout using Equations 9, 10, and 11,

$$t_{aes} = Dropout(W_{aes}h_{aes} + b_{aes}), \tag{9}$$

$$t_{spe} = Dropout\big(W_{spe}h_{spe} + b_{spe}\big), \tag{10}$$

$$t_{sem} = Dropout(W_{sem}h_{sem} + b_{sem}), \tag{11}$$

with $t, \in \mathbb{R}^D$ and $D = 256$.

The LAB channel components were not passed through the Vision Transformer primarily because the ViT pretraining itself aims to optimize the RGB semantics. This may affect the pretrained semantic consistency. Moreover, the current model maintains a balance between the feasibility of deployment and representational complementarity. However, from an architectural perspective, a model exploring the benefits of a fully transformer-based spectral pathway certainly holds potential.

3.2.2 Multi-head Self-Attention Fusion

The three projected features were stacked as a sequence of tokens using Equation 12,

$$T = \begin{bmatrix} t_{aes} \\ t_{spe} \\ t_{sem} \end{bmatrix} \in \mathbb{R}^{3 \times D} \tag{12}$$

and passed through a multi-head self-attention layer with $H = 4$ heads. For each head h, query, key, and value metrics were computed using Equation 13,

$$Q_h = TW_h^Q, K_h = TW_h^K, V_h = TW_h^V, \tag{13}$$

where $d_k = D/H$ is the per-head dimension. The outputs from all heads were concatenated and projected using Equation 14,

$$MHA(T) = [Attn_1(T) \| \ldots \| Attn_H(T)]W^O. \tag{14}$$

A dropout layer was applied to obtain the attended tokens $\hat{T} \in \mathbb{R}^{3 \times D}$.

3.2.3 Gated Branch Fusion

The attended tokens were then fused via a learnable gating mechanism that dynamically weights the contribution of each branch for every sample using Equation 15. Let,

$$\hat{T} = \begin{bmatrix} \hat{t}_{aes} \\ \hat{t}_{spe} \\ \hat{t}_{sem} \end{bmatrix} \hat{t}_i \in \mathbb{R}^D. \tag{15}$$

The tokens were flattened into a single vector of $u = vec(\hat{T}) \in \mathbb{R}^{3D}$, used as input to a fully connected gating layer $g = W_g u + b_g \in \mathbb{R}^3, \alpha = Softmax(g)$, where $\alpha = (\alpha_{aes}, \alpha_{spe,} \alpha_{sem})$ are non-negative branch weights satisfying $\sum_i \alpha_i = 1$. The fused representation was computed as a convex combination using Equation 16,

$$f = \sum_{i \in \{aes,spe,sem\}} \alpha_i \hat{t}_i \in \mathbb{R}^D \tag{16}$$

This made it possible to depend more on the LAB branch in cases where color features were more prevalent, and vice versa on the semantic branch of the ViT model. In addition, it was ensured that the three branches encode complementary, not redundant, features. The alignment of the projection and the attention weighting scheme reduces redundancy and allows the network to concentrate on the most prevalent modalities of the inputs. Furthermore, issues

of collapse in the gates were overcome using SoftMax normalization, regularization, and supervised contrastive alignment.

The fusion mechanism allows the network to dynamically choose which modality has more information about each image. When the dominant contamination is discoloration, the LAB feature receives more attention. When the degradation of the texture is more emphasized, the influence of the RGB feature became more important. For images that are ambigiously spoiled, the semantic branch of ViT helps disambiguate.

3.2.4 Classification and Projection Heads

The fused representation f was passed through a lightweight classifier composed of a fully connected layer with ReLU activation and dropout, followed by the final layer as in Equations 17 and 18,

$$h = ReLU(W_1 f + b_1), \tag{17}$$

$$o = W_2 h + b_2 \in \mathbb{R}^C, \tag{18}$$

where $C = 8$ is the number of spoilage classes. The predicted class probabilities were obtained using the SoftMax function as illustrated in Equation 19,

$$P_c = \frac{\exp(O_c)}{\sum_{k=1}^{C} \exp(O_k)} \tag{19}$$

In parallel, a projection head $g(\cdot)$ mapped f into a lower-dimensional embedding $z \in \mathbb{R}^{d_{proj}}$ for supervised contrastive learning as shown in Equation 20,

$$z = g(f) = W_4 ReLU(W_3 f + b_3) + b_4, \qquad d_{proj} = 128 \tag{20}$$

3.3 Training Objective and Optimization

3.3.1 Class Weighting and Label

Class weights w_c were computed as inverse frequencies over the training set to mitigate class imbalance across the eight categories of the image datasets. Equation 21 illustrates the computation as let n_c denote the number of training samples belonging to class c. The unnormalized inverse frequency was $1/n_c$, these values were rescaled to sum to the number of classes.

$$w_c = \frac{\frac{1}{n_c}}{\sum_{k=1}^{C} \frac{1}{n_k}} \cdot C \tag{21}$$

Moreover, to regularize the classifier and reduce overconfidence, label smoothing with factor ε was applied. For the ground-truth class y of a sample, the smoothed target distribution $\tilde{y}$ was defined as in Equation 22,

$$\tilde{y}_c = \begin{cases} 1 - \varepsilon, & c = y, \\ \dfrac{\varepsilon}{C-1}, & c \neq y \end{cases} \tag{22}$$

The class-weighted and label-smoothed cross-entropy loss for a sample was computed using Equation 23,

$$L_{CE}(p, \tilde{y}) = -\sum_{c=1}^{C} w_c \tilde{y}_c \log p_c \tag{23}$$

3.3.2 MixUp and CutMix Augmentation

To further improve generalization under limited data, the model was trained with a combination of MixUp and CutMix applied jointly on the RGB and LAB inputs. For a mini-batch of size B, a random permutation of indices was sampled, and each pair $(x_i,\ y_i)$ and $(x_j,\ y_j)$ was combined. In the MixUp case, inputs and labels were linearly interpolated using $\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \tilde{y} = y_i + (1 - \lambda)y_j, \lambda \sim Beta(\alpha, \alpha)$. Since labels are represented as integers, the implementation kept two label tensors y_a and y_b and the mixing coefficient λ. The cross-entropy loss under MixUp was computed using Equation 24,

$$L_{Mix} = \lambda L_{CE}(p, y_a) + (1 - \lambda)\, L_{CE}(p, y_b) \tag{24}$$

In the CutMix case, a random rectangular region was cut from x_i and replaced by the corresponding region from x_j. If the patch covered area $\lceil R \rceil$ and the full image area was HW, the effective mixing coefficient was computed using Equation 25,

$$\lambda = 1 - \frac{|R|}{HW} \tag{25}$$

The same mixed loss formulation as above was used, and, in the implementation, each batch was randomly assigned to no augmentation, MixUp, or CutMix with equal probability, which provides a diverse set of training perturbations.

3.3.3 Supervised Contrastive Loss

In addition to cross-entropy, a supervised contrastive loss was imposed on the projection embeddings z to encourage samples from the same class to cluster together in feature space. For a batch of B embeddings z_t and labels y_i, the embeddings were first L2-normalized as illustrated in Equation 26. On the other hand, pairwise similarities were computed using the dot product scaled by a temperature parameter τ as shown in Equation 27. For an anchor sample i, the set of positive indices was $P(i) = \{j \neq i : y_j = y_i\}$. The supervised contrastive loss [16] for i was computed using Equation 28, and the batch loss was the mean over all anchors using Equation 29.

$$\hat{z}_i = \frac{z_i}{\|z_i\|_2} \tag{26}$$

$$s_{ij} = \frac{\hat{z}_i^{\mathrm{T}}\, \hat{z}_j}{\tau} \tag{27}$$

$$\ell_i = \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s_{ip})}{\sum_{a \neq i} \exp(s_{ia})} \tag{28}$$

$$L_{SupCon} = \frac{1}{B}\sum_{i=1}^{B} \ell_i \tag{29}$$

3.3.4 Total Loss and Optimization

During the training, the model received two forward passes per batch, an unmixed pass for the contrastive loss and a mixed pass for the cross-entropy loss under the MixUp and CutMix. The parameters of all branches and fusion layers were optimized using Adam with a learning rate of $\eta = 10^{-4}$ and a weight decay of 10^{-4}, and cosine annealing scheduler gradually decreased the learning rate over 30 epochs.

3.3.5 Training Algorithm

Algorithm 1: DASS-Net Hybrid Training Protocol

Input:
Dataset D = { (x_i, y_i) }_{i=1..N} # images x, labels y
Hyperparameters:
  epochs = E
  batch_size = B
  lr (initial learning rate)
  weight_decay
  IMG_SIZE, RANDOM_SEED
  use_mixup_cutmix (bool), alpha_mix, alpha_cut
  supcon_temperature, supcon_lambda
Model components:
  model θ = DASSNetHybrid3Branch(...) # returns (logits, fused, z_proj)
  optimizer = AdamW(θ, lr, weight_decay)
  scheduler = CosineAnnealingLR(optimizer, T_max=E)
  criterion = CrossEntropyLoss(weight=class_weights, label_smoothing=λ_smooth)
Output:
Trained model θ* (parameters saved at best validation accuracy)
History H with per-epoch metrics (train/val loss, train/val acc, lr)
Helper functions:
Mixup/CutMix(rgb, lab, targets, α_mix, α_cut) -> (rgb_m, lab_m, y_a, y_b, λ, aug_type)
MixupCriterion(criterion, preds, y_a, y_b, λ) -> mixed_ce_loss
SupervisedContrastiveLoss(z, labels, temperature) -> con_loss
Evaluate(model, loader) -> (loss, acc) # no-grad eval
Procedure:
1. Initialize random seeds and device (CPU/GPU).
2. Build dataset splits: train_loader, val_loader, test_loader.
3. Compute class_weights from train set and move to device.
4. Initialize model θ on device, optimizer, scheduler, criterion.
5. best_val_acc ← 0.0
6. For epoch = 1 to E:
  6.1 Set model to train mode.
  6.2 Initialize accumulators: train_loss=0, train_correct=0, total=0.
  6.3 For each batch (rgb, lab, labels) in train_loader:
    - Move data to device.

- optimizer.zero_grad()
- Forward clean: logits_clean, fused_clean, z_clean = model(rgb, lab)
- Apply data-mix augmentation:
  rgb_m, lab_m, y_a, y_b, λ, aug_type = Mixup/CutMix(rgb, lab, labels)
  Forward mixed: logits_mixed, _, _ = model(rgb_m, lab_m)
- Compute classification loss:
  if λ == 1.0:
    ce_loss = criterion(logits_mixed, y_a)
  else:
    ce_loss = MixupCriterion(criterion, logits_mixed, y_a, y_b, λ)
- Compute contrastive loss on clean embeddings:
  con_loss = SupervisedContrastiveLoss(z_clean, labels, supcon_temperature)
- Total batch loss:
  loss = ce_loss + supcon_lambda * con_loss
- Backpropagate: loss.backward(); optimizer.step()
- Update running counters using logits_clean predictions:
  preds = argmax(logits_clean, dim=1)
  train_loss += loss.item() * batch_size
  train_correct += sum(preds == labels)
  total += batch_size

6.4 Compute epoch-level train_loss /= total; train_acc = train_correct / total
6.5 Set model to eval mode.
6.6 Evaluate on val_loader without mix augmentation:
val_loss, val_acc = Evaluate(model, val_loader)
6.7 scheduler.step(); record lr
6.8 Save epoch metrics to H.
6.9 If val_acc > best_val_acc:
best_val_acc ← val_acc
save model weights θ* = θ (checkpoint)
7. End For
8. Load θ* (best model), then compute final test metrics: test_acc, f1, confusion matrix, per-class reports, Grad-CAM, failure case visualizations, ablation tables, cross-validation results.

The training procedure involved a two-step per epoch process: (i) a training step, in which the model parameters were optimized using a combined classification loss and supervised contrastive loss, and (ii) an evaluation step used for model selection. During the training procedure, the model essentially went through the mini-batches twice conceptually: (i) first, to compute the clean embeddings required by the supervised contrastive loss, and (ii) second, by traversing the augmented data to compute the cross-entropy loss and a scaled supervised contrastive loss. Additionally, during this specific setup, the data augmentation is considered probabilistic, and the contribution from MixUp/CutMix is addressed by using the correct mixed-label loss. Following every epoch, the model is evaluated on the clean validation set without mixing to compute the validation loss and accuracy. The learning rate is then updated using a cosine annealing scheduler within each epoch, and the model parameters that performed best on the validation accuracy will be checkpointed for final testing.

3.4 Evaluation Protocol

3.4.1 Performance Metrics

After training, the model was evaluated on the hold-out test set. For each test image, the predicted class $\hat{y}$ was obtained as the index of the maximum logit. Overall accuracy was defined as in Equation 30,

$$Accuracy = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} 1[\hat{y}_i = y_i] \tag{30}$$

where $1[\cdot]$ is the indicator function. In addition, per-class precision, recall, and F1-score were computed from the confusion counts. For every class, true positives (TP), false positives (FP), and false negatives (FN) were derived from the confusion matrix. The metrics were then computed using Equation 31, Equation 32, and Equation 33.

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{31}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{32}$$

$$F1_c = \frac{2Precision_c Recall_c}{Precision_c + Recall_c} \tag{33}$$

Macro-F1 was obtained as the unweighted average of $F1_c$ across classes, while weighted-F1 weighted each $F1_c$ by its class support. Moreover, the confusion matrix was also used to summarize the number of predictions in each true-predicted pair. The row-normalized version was also reported by dividing each row by its sum.

3.4.2 Safety-Critical Error Analysis

Considering that misclassifying spoiled food as fresh has more severe consequences than the reverse, a safety-oriented metric was defined. Let indices $0-3$ denote fresh classes and $4-7$ spoiled classes; the total number of spoiled samples in the test set was defined as in Equation 34. The number of spoiled items incorrectly predicted as fresh was computed using Equation 35, and the spoiled-as-fresh rate was then computed using Equation 36. This metric directly quantifies the risk of unsafe recommendations in hospitality operations.

$$N_{spoiled} = \sum_{i=4}^{7} \sum_{j=0}^{7} C_{ij} \tag{34}$$

$$N_{spoiled \rightarrow fresh} = \sum_{i=4}^{7} \sum_{j=0}^{7} C_{ij} \tag{35}$$

$$r_{spoiled \rightarrow fresh} = \frac{N_{spoiled \rightarrow fresh}}{N_{spoiled}} \tag{36}$$

3.4.3 Ablation Experiments

Since the impact of incorrectly labeling spoiled food as fresh food would result in more severe outcomes compared to the opposite action, a safety-minded measure was developed. Where the indices from 0-3 indicated the fresh categories and from 4-7 indicated the spoiled categories, the total number of spoiled examples from the testing set was stated in Equation 34. The number of spoiled items labeled incorrectly as fresh food was stated in Equation 35; then the rate of spoiled items labeled as fresh food was stated in Equation 36.

3.4.4 Explainability and Failure-Case Analysis

In this study, qualitative analysis was performed using gradient-weighted class activation maps (Grad-CAM) on both the RGB and LAB branches to visualize the spatial regions that contributed most to spoilage predictions. For each selected image and target class c_1 Grad-CAM computed the importance weights α_k^c of feature maps A^k in the last convolutional layer and a generated heatmap using Equation 37,

$$L_{Grad-CAM}^{c} = ReLU\left(\sum_{k} \alpha_k^c A^k\right) \tag{37}$$

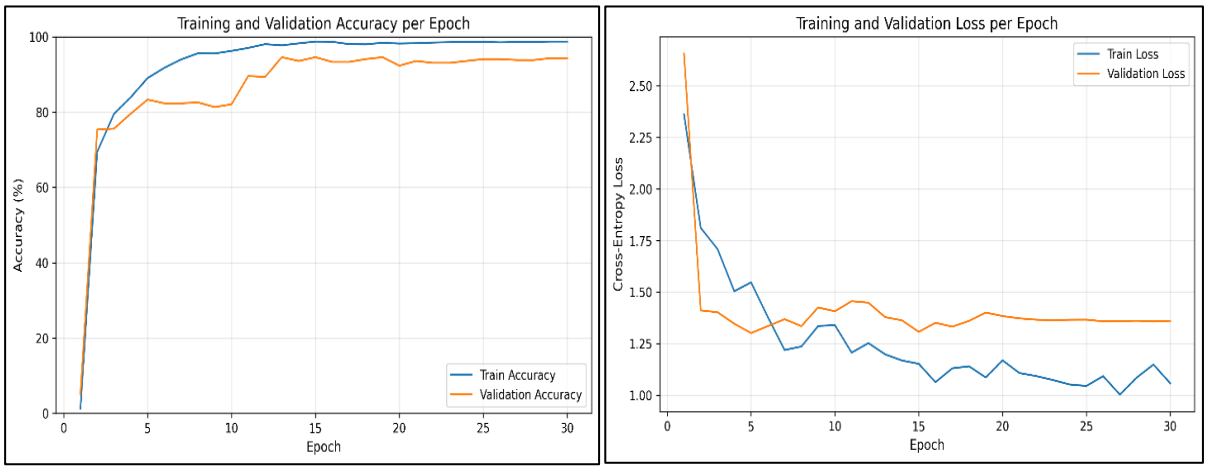
which was upsampled and overlaid on the input image. These visualizations highlighted mold spots, discoloration, and texture changes that the model relied on. In addition, misclassified cases were categorized into fresh-to-spoiled and spoiled-to-fresh errors and inspected manually to understand systematic failure modes, such as naturally browned but safe items or very subtle spoilage patterns.

4. Results and Discussion

The final section of this paper provides the outcome of the empirical assessment of the proposed hybrid DASS-Net model when applied to image-based classification problems in food spoilage. It includes the model's performance during training, the ablation study, performance on each class, confusion matrices, explanation visuals, and failure case analyses. The proposed model was tested using an open-source dataset retrieved from the official Kaggle platform [17] for eight different classes that involved fresh and spoiled products of bread, dairy products, fruits, and vegetables. The entire experimental setup was carried out by employing Python version 3.10.5 in the Jupyter Notebook environment, along with an NVIDIA GeForce RTX 3050 graphics card, an AMD Ryzen 5 processor, and 16 GB of RAM.

4.1 Training and Validation Performance

Figure 2(a) depicts the training and validation accuracy over the epochs. The accuracy of the validation data for the Hybrid DASS-Net model rose from 65.17% in Epoch 1 to 94.33% in the final epoch, thereby ensuring a smooth convergence process. As the optimization was smooth with no signs of overfitting, the training accuracy began to increase. Figure 2(b) depicts the loss for training and validation. As both the training and validation loss reach convergence at around Epoch 15, it is evident that the stability in accuracy is appropriate, thereby avoiding overfitting. According to a previous study, in the context of analyzing food images, using contrast regularization helps ensure the compactness of features and separation of classes [18].



(a) (b)

Figure 2. Training and Validation Performance of the Proposed DASS-Net Model. (a) is the Training and Validation Accuracy, and (b) is the Training and Validation Loss

4.2 Ablation Study

The ablation study compares three model variants: (i) RGB-Only ResNet-18, (ii) LAB-Only ResNet-18, and (iii) the proposed Hybrid DASS-Net.

4.2.1 Quantitative Ablation Metrics

Table 2. Ablation Performance Comparison Across RGB-Only, LAB-Only, and Hybrid DASS-Net

Model	Val Accuracy	Macro-F1	Weighted F1	Spoiled-as-fresh rate
RGB-Only ResNet18	71.15	70.80	70.75	28.44
LAB-Only ResNet18	68.50	68.41	68.48	30.28
Hybrid DASS-Net	94.33	94.00	94.00	13.30

Table 2 summarizes the performance of three model variants based on their validation accuracy, macro F1-score, weighted F1-score, and spoiled-as-fresh error rate. Based on the results, the Hybrid DASS-Net tended to outperform both baseline models in every metric. With a weighted F1 of 94%, and a macro F1-score of 94%, the hybrid DASS-Net model generated the lowest safety-critical error rate of 13.30%. Moreover, this improvement also highlights the advantages of representation fusion across various dimensions, from spectral (LAB) to semantic (ViT), and further to aesthetic (RGB). This is supported by previous studies indicating that RGB frequently fails to detect minute discoloration signs that are captured by the LAB color space. Additionally, substantial global context awareness provided by the Vision Transformers generated improved spoiled detection when both are combined.

From the ablation results, the removal of individual elements of the framework tends to result in performance degradation, suggesting that the proposed branches and learning elements are complementary rather than mere duplications of already existing capacities. Consistent with

the Grad-CAM and interpretive analyses, it follows that the contribution of the RGB pathway tends to be sensitive to the visible surface, that the LAB pathway facilitates the understanding of the spectra, and that the ViT pathways help in the case of visible ambiguities. While the current scope does not allow for a full factor-isolation analysis, further research may explore larger-scale controlled ablation on larger datasets for even more refined contribution attribution.

4.3 Per-Class Evaluation Metrics

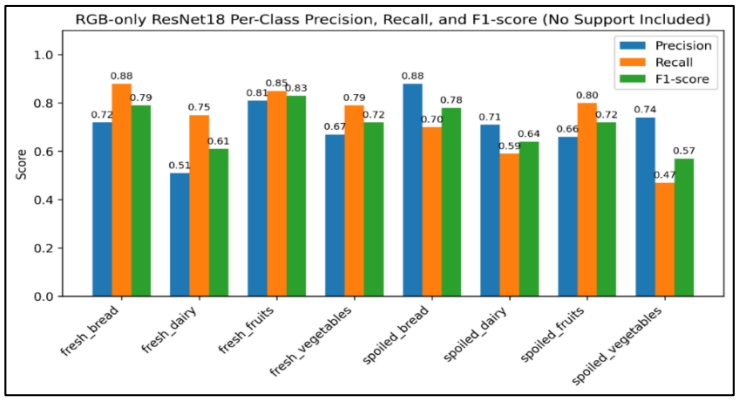


Figure 3. Per-Class Precision, Recall, and F1-Score of the RGB-only Model Variant

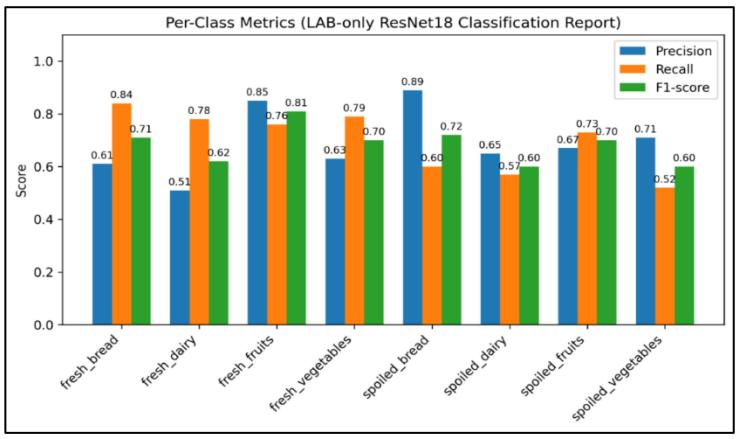


Figure 4. Per-Class Precision, Recall, and F1-Score of the LAB-only Model Variant

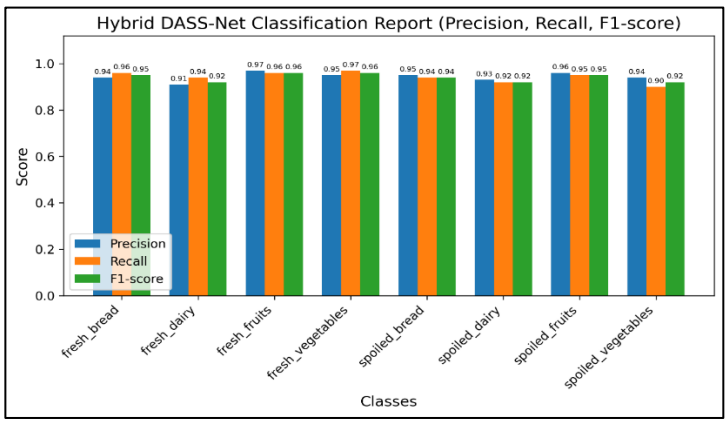


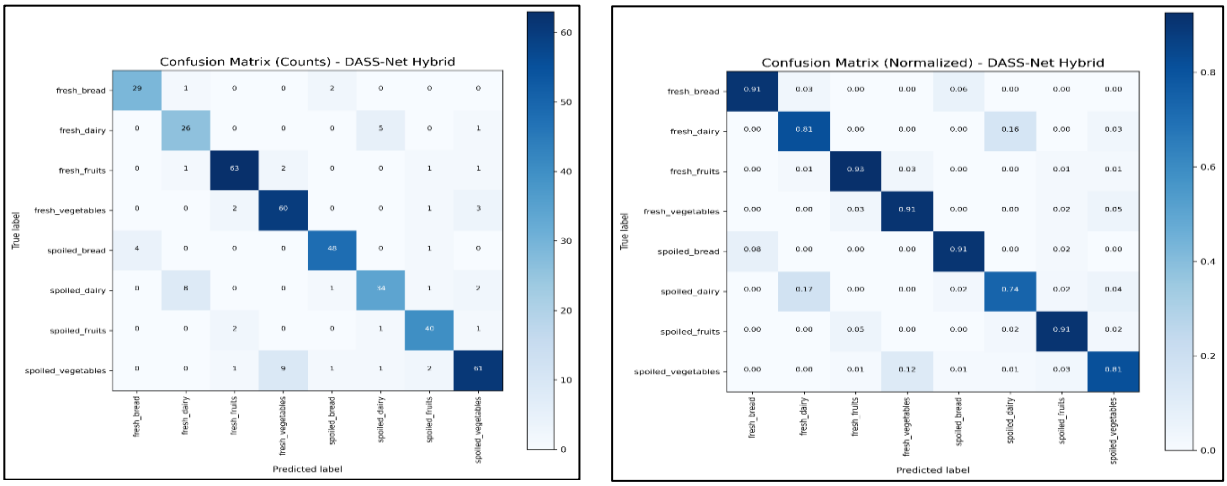
Figure 5. Per-class Precision, Recall, and F1-Score of the Proposed DASS-Net Model

Figure 3 above depicts the precision, recall, and F1-score for each class based on the RGB-Only model variant. From Figure 3, it can be seen that the model has a very low F1-score of 61% and 57% for both fresh dairy products and vegetables, respectively, in detecting spoilage. This serves as an indication that it is not very straightforward or easy to identify discoloration through observation of the RGB channel alone in both dairy products and vegetables.

Figure 4 below highlights the precision, recall, and F1-score of each class in the LAB-Only model variant. Based on the obtained results, it is safe to state that the LAB-Only branch did not perform well with the classes containing dairy products relative to the fruits and vegetables. The assertion that the LAB space enhances spectral sensitivity but lacks semantic understanding is indeed valid, as it is demonstrated in the transformer.

Figure 5 above shows the precision, recall, and F1 score per class attained using the proposed hybrid DASS-Net model. Clearly, the DASS-Net has recorded the highest scores in all classes. Spoiled classes have also attained excellent F1-scores, including 94% for spoiled bread, 92% for spoiled dairy, 95% for spoiled fruits, and 92% for spoiled vegetables. This indicates that the multi-branch fusion has managed to overcome the ambiguity in the classes, thereby improving the identification of superficial defects.

4.4 Confusion Matrix



(a) (b)

Figure 6. Confusion Matrix of the Proposed Hybrid DASS-Net Model

Figure 6 depicts the confusion matrix of the proposed hybrid model, DASS-Net. The findings indicate that the model was able to accurately categorize most of the images in both fresh and spoiled categories without much confusion. It can be seen from Figure 6a above that in the counts and fresh bread categorization, the model accurately categorized 29 images as fresh, while only 2 were confused with fresh dairy and fresh fruits. However, unlike in counts and fresh bread, fresh dairy showed moderate confusion in the spoiled dairy, spoiled fruits, and spoiled vegetables categories. In the spoiled categories, the model was able to accurately categorize more than 60 images, while some were confused in their prediction, particularly in spoiled dairy, which was confused with the fresh dairy category. Most of the images were along the diagonal, which explains the model's performance. The model was able to accurately categorize most images in the spoiled categories, particularly in spoiled bread (45), spoiled

dairy (48), spoiled fruits (40), and spoiled vegetables (63). However, some images were confused in the spoiled dairy category, which was classified into the fresh dairy category.

On the other hand, the normalized confusion matrix is presented in Figure 6(b). The results shows high precision and recall scores of <90% for fresh bread, fresh fruits, spoiled fruits, and spoiled vegetables. There is moderate confusion in dairy classes for fresh dairy to spoiled dairy (0.16) and spoiled dairy to fresh dairy (0.17). The results imply that dairy-related foods are the most visually ambiguous category. According to a previous study [21], it is hard to detect early-stage spoilage of dairy-related foods only by vision as, they usually present subtle discoloration and/or texture changes.

The results confirm the effectiveness of multi-branch fusion, in accordance with reports that the combination of RGB-spectral-semantic modalities improves food-quality prediction tasks [22]. Another finding is that dairy, whether fresh or spoiled, is the most difficult class to recognize and is misclassified at rates between 10 and 17% in both matrices. According to Nur et al. (2024), early spoilage of dairy products often involves micro-texture changes or very slight discoloration that may not be strongly reflected in RGB. Moreover, cheese and milk-based foods bring intra-class confusion because they naturally vary in hue, and dairy surfaces can reflect light irregularly [24], thus causing modeling difficulties in both RGB and LAB spaces.

This problem is very meaningful to hotel and restaurant operations, as serving spoiled dairy can lead to gastrointestinal illness. Therefore, misclassifying some spoiled items as fresh, particularly in dairy, is a major concern. Even a small misclassification rate may imply high operational risk in food-safety contexts. Additionally, minimizing false negatives (spoiled to fresh) is more important than maximizing accuracy [25]. Even though it is balanced, reality has cases of spoilage rarity. To mitigate this problem, it is necessary to introduce what is referred to as "spoiled-as-fresh" safety, which will provide valuable information on high-risk, rather than accuracy, cases of spoilage.

4.5 Grad-CAM Visualization

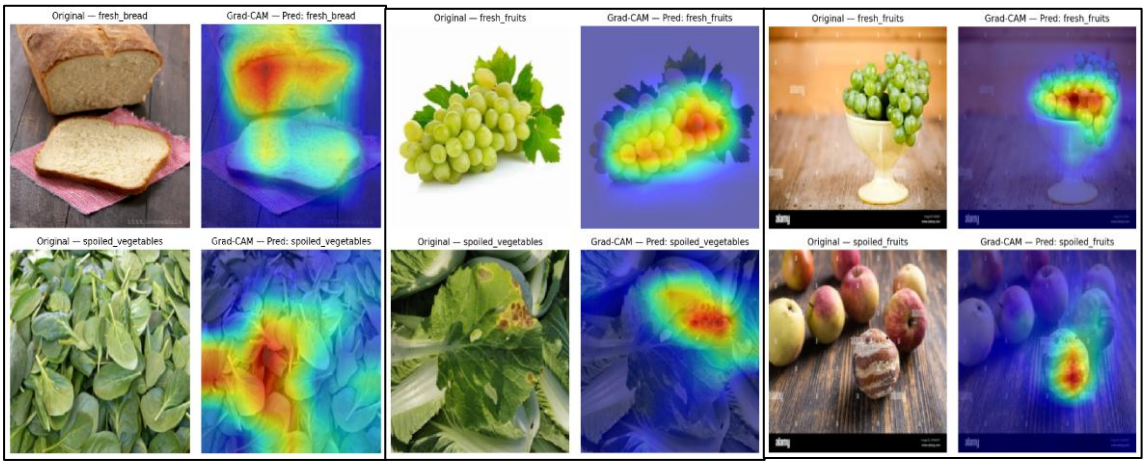


Figure 7. RGB-Branch Grad-CAM Visualization

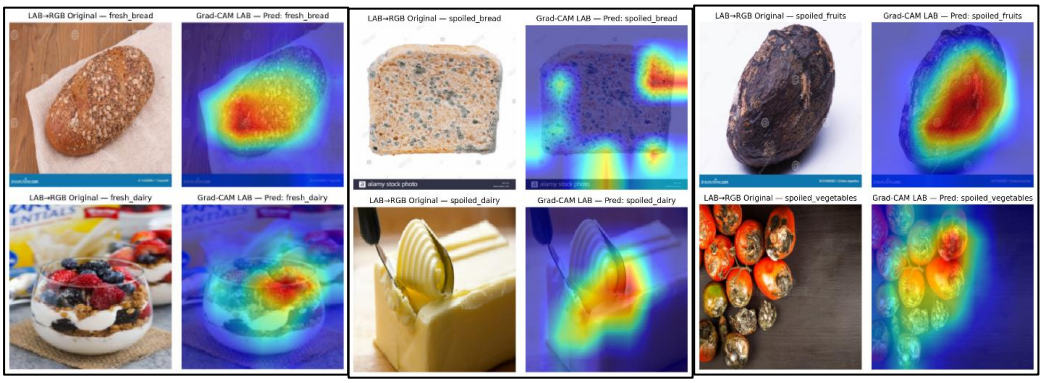


Figure 8. LAB-Branch Grad-CAM Visualization

Figures 7 and 8 show the results obtained from the RGB and LAB branches regarding attribute interpretation in both fine-tuned Grad-CAM models. As observed from Figure 7, the Grad-CAM model in the RGB branch emphasized defects such as texture degradation, spots from mold growth, and some irregularities in food surfaces, whereas in the LAB branch Grad-CAM, emphasized discolored regions due to early browning or chromatic redistribution that are not visible in the original images in RGB domain. Activation maps from both multiple branch approaches justify that each branch has provided complementary information through unique activation maps. This supports the theoretical strength in fusing multiple features in one [26]. Moreover, both targeted areas in images could benefit from global attention in the ViT architecture, which has also assisted in recognizing BC in boundary-sensitive regions, as well as classifying them into either spoilage or natural variations in color in food images. Experiments demonstrate that ViTs are effective in non-local interactions between features in tasks such as image processing [27], which is attributed to better performance over purely CNN-based architectures in our project.

4.6 Failure Case Analysis

Figure 9 highlights samples that were misclassified as being spoiled or fresh. The number of spoiled samples predicted to be fresh was 32, and the number of fresh samples predicted to be spoiled was 39. The misrepresentation of spoiled samples as fresh samples usually occurs when the growth rate of mold is low. This is combined with lighting that suppresses the appearance of color. The samples that were misclassified as spoiled likely contained strong shadows and glare from the cameras, resulting in false color degradation. In contrast, the samples misclassified as fresh were those that contained natural variations in color. This was expected, as natural variation makes visual inspection difficult. The low percentage (13.30%) rate of misrepresentation from the spoiled to the fresh class proves that the model is reliable.

The identified failure instances have also provided insights into the performance of the tested model under challenging visual conditions. As previously mentioned, some instances of the "spoiled-as-fresh" failure occurred when discoloration visibility was masked by lighting conditions, causing only partial surface discoloration to be noticed. The "fresh-as-spoiled" failure instances related to shadow and glare were linked to surface color ambiguities. To complement the identified instances, insights from the Grad-CAM tasks highlighted concerns with the LAB approach in understanding the spectral discoloration pattern, while the approach

of the ViT block focused on comprehending the entire context of challenging visual instances. Although the study does not examine the weighted dynamics of the fusion approach quantitatively, it indicates that the approach might be beneficial across different modalities under distinct lighting conditions.

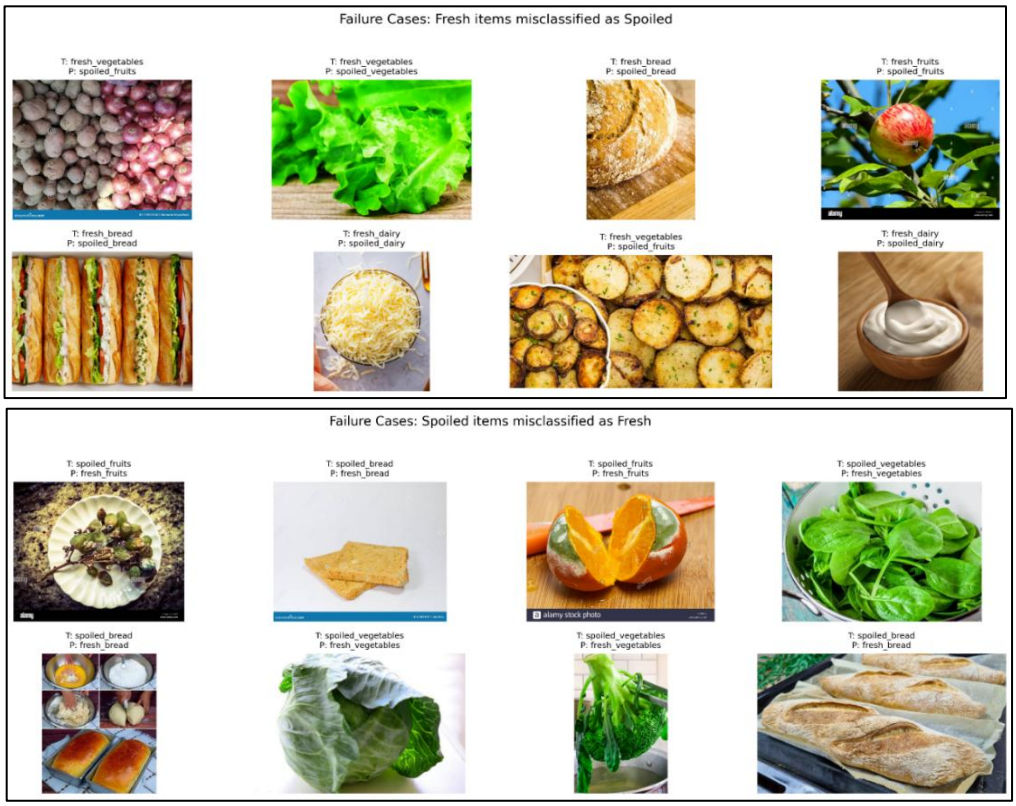


Figure 9. Failure Cases of DASS-Net (Misclassified Samples)

In this particular study, the procedure of color conversion and normalization in the LAB color space functioned properly in a real-world setting, and no instability issues in performance calculations were observed during the experiments. However, a specific sensitivity analysis of the normalization procedure in relation to calibration may be of significant interest from a future perspective.

5. Conclusion

The objective of the study was the development of the DASS-Net Multi-Branch Neural Net Approach. The neural net would be able to classify images of food spoilage into the eight categories used in the spoiled and fresh image datasets. The study concluded that the approach led to an accuracy of 94.33% for the validation set. The approach was better compared to the single-stream approaches regarding accuracy and safety. The study has relevance because it introduced an adaptive approach to fusion that implemented multiple features of the models. The study presented new approaches that could be applied during the development of the neural nets, including the application of the SupCon approach. The study has played a significant role because it demonstrated the application of safety-critical approaches, specifically the dual Grad-CAM. The study was able to consider the rate of spoilage of food products once they were fresh. The study also has practical implications because it deals with the application of smart kitchen systems and hotels. For future research, both multispectral and hyperspectral

imaging can be combined to indicate biochemical indicators of spoilage not apparent in either RGB or LAB images. The proposed system can be expanded to apply real-time video monitoring involving light transformer architecture systems. Additionally, more varieties of food could be introduced to enhance robustness across various hospitality settings.

References

[1] Elimelech, Efrat, Eyal Ert, Yael Parag, and Guy Hochman. "Exploring The Impact of Visual Perception and Taste Experience on Consumers' Acceptance of Suboptimal Fresh Produce." Sustainability 16, no. 7 (2024): 2698.

[2] Okpala, Charles Odilichukwu R., and Małgorzata Korzeniowska. "Understanding the Relevance of Quality Management in Agro-Food Product Industry: From Ethical Considerations to Assuring Food Hygiene Quality Safety Standards and Its Associated Processes." Food Reviews International 39, no. 4 (2023): 1879-1952.

[3] Creed, P. G. "Sensory Quality Control in Foodservice." Sensory Analysis for Food and Beverage Quality Control (2010): 316-336.

[4] Cohen Hakmon, May, Keren Buhnik-Rosenblau, Hila Hanani, Hila Korach-Rechtman, Dagan Mor, Erez Etkin, and Yechezkel Kashi. "Early Detection of Food Safety and Spoilage Incidents Based on Live Microbiome Profiling and PMA-qPCR Monitoring of Indicators." Foods 13, no. 15 (2024): 2459.

[5] Segbedzi, Cynthia Esinam, Edward Wilson Ansah, and Daniel Apaak. "Compliance to Food Safety Standards: Determining the Barriers Within the Hotel Industry." PLOS Global Public Health 5, no. 11 (2025): e0002771.

[6] Balakrishnan, P., A. Anny Leema, N. Jothiaruna, Purshottam J. Assudani, K. Sankar, Madhusudan B. Kulkarni, and Manish Bhaiyya. "Artificial Intelligence for Food Safety: From Predictive Models to Real-World Safeguards." Trends in Food Science & Technology 163 (2025): 105153.

[7] Ahmad, Muhammad, Salvatore Distefano, Adil Mehmood Khan, Manuel Mazzara, Chenyu Li, Hao Li, Jagannath Aryal, Yao Ding, Gemine Vivone, and Danfeng Hong. "A Comprehensive Survey for Hyperspectral Image Classification: The Evolution from Conventional to Transformers and Mamba Models." Neurocomputing (2025): 130428.

[8] Anwar, Abbas, Saira Kanwal, Muhammad Tahir, Muhammad Saqib, Muhammad Uzair, Mohammad Khalid Imam Rahmani, and Habib Ullah. "Image Aesthetic Assessment: A Comparative Study of Hand-Crafted & Deep Learning Models." IEEE Access 10 (2022): 101770-101789.

[9] Pu, Yumei, Danfei Liu, Siyuan Chen, and Yunfei Zhong. "Research Progress on the Aesthetic Quality Assessment of Complex Layout Images Based on Deep Learning." Applied Sciences 13, no. 17 (2023): 9763.

[10] Treepong, Panisa, and Nawanol Theera-Ampornpunt. "Early Bread Mold Detection Through Microscopic Images Using Convolutional Neural Network." Current Research in Food Science 7 (2023): 100574.

[11] Jubayer, Fahad, Janibul Alam Soeb, Abu Naser Mojumder, Mitun Kanti Paul, Pranta Barua, Shahidullah Kayshar, Syeda Sabrina Akter, Mizanur Rahman, and Amirul Islam. "Detection of Mold on the Food Surface Using YOLOv5." Current Research in Food Science 4 (2021): 724-728.

[12] Feng, Chao-Hui. "Colour Analysis of Sausages Stuffed with Modified Casings Added with Citrus Peel Extracts Using Hyperspectral Imaging Combined with Multivariate Analysis." Sustainability 16, no. 19 (2024): 8683.

[13] Cheng, Lijuan, Guishan Liu, Jianguo He, Guoling Wan, Chao Ma, Jingjing Ban, and Limin Ma. "Non-Destructive Assessment of the Myoglobin Content of Tan Sheep Using Hyperspectral Imaging." Meat Science 167 (2020): 107988.

[14] Foong, Chai C., Goh K. Meng, and Lim L. Tze. "Convolutional Neural Network Based Rotten Fruit Detection Using Resnet50." In 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), IEEE, 2021, 75-80.

[15] Gillespie, James, Jordan Vincent, Omar Dib, Matthias Heiden, and Joan Condell. "Handheld Spectroscopy for Dairy Quality Control: Towards a Real Time Milk Spoilage Classification Model." In International Conference on Ubiquitous Computing and Ambient Intelligence, Cham: Springer Nature Switzerland, 2024, 829-840.

[16] Khosla, Prannay, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. "Supervised Contrastive Learning." Advances in neural information processing systems 33 (2020): 18661-18673.

[17] Fresh and Spoiled Food Image Dataset. (n.d.). Retrieved November 25, 2025, from https://www.kaggle.com/datasets/maheen00shahid/fresh-and-spoiled-food-image-dataset

[18] Govindarajan, Vijay, and Junaid Hussain Muzamal. "Advanced Cloud Intrusion Detection Framework Using Graph Based Features Transformers and Contrastive Learning." Scientific Reports 15, no. 1 (2025): 20511.

[19] Chen, Dongyu, and Haitao Zhao. "CCD-Net: Color-Correction Network Based on Dual-Branch Fusion of Different Color Spaces for Image Dehazing." Applied Sciences 15, no. 6 (2025): 3191.

[20] Hatamizadeh, Ali, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. "Global Context Vision Transformers." In International Conference on Machine Learning, pp. 12633-12646. PMLR, 2023.

[21] Salehi, Fakhreddin. "Quality, Physicochemical, And Textural Properties of Dairy Products Containing Fruits and Vegetables: A Review." Food Science & Nutrition 9, no. 8 (2021): 4666-4686.

[22] Lun, Zhichen, Xiaohong Wu, Jiajun Dong, and Bin Wu. "Deep Learning-Enhanced Spectroscopic Technologies for Food Quality Assessment: Convergence and Emerging Frontiers." Foods 14, no. 13 (2025): 2350.

[23] Nur Farzanah Faghira Kamarudin, Puteri, Nik Mohd Zarifie Hashim, Masrullizam Mat Ibrahim, and Mahmud Dwi Sulistiyo. "Milk Spoilage Classification Through Integration

of RGB and Thermal Data Analysis." International Journal of Computing and Digital Systems 15, no. 1 (2024): 1839-1851.

[24] Kilcawley, Kieran N., Hope Faulkner, Holly J. Clarke, Maurice G. O'Sullivan, and Joseph P. Kerry. "Factors Influencing the Flavour of Bovine Milk and Cheese from Grass Based Versus Non-Grass Based Milk Production Systems." Foods 7, no. 3 (2018): 37.

[25] Practices of Science: False Positives and False Negatives | manoa.hawaii.edu/ExploringOurFluidEarth. (n.d.). Retrieved December 9, 2025, from https://manoa.hawaii.edu/exploringourfluidearth/chemical/matter/properties-matter/practices-science-false-positives-and-false-negatives

[26] Han, Xiaofeng, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang et al. "Multimodal Fusion and Vision-Language Models: A survey for robot vision." Information Fusion (2025): 103652.

[27] Maurício, José, Inês Domingues, and Jorge Bernardino. "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review." Applied Sciences 13, no. 9 (2023): 5521.

[28] Yang, Yican, Nuwan K. Wijewardane, Lorin Harvey, and Xin Zhang. "Beyond Flips and Rotations: Evaluating NeRF and 3DGS-Based Synthetic View Augmentation: A Case Study of Detecting Visual Skinning Damage on Sweetpotato Storage Roots." Journal of Agriculture and Food Research (2025): 102517.

[29] Elkins, Andrew, Felipe F. Freitas, and Verónica Sanz. "Developing an App to Interpret Chest X-Rays to Support the Diagnosis of Respiratory Pathology with Artificial Intelligence." arXiv preprint arXiv:1906.11282 (2019).

[30] Araujo, Sufola Das Chagas Silva, Goh Kah Ong Michael, Uttam U. Deshpande, Sudhindra Deshpande, Manjunath G. Avalappa, Yash Amasi, Sumit Patil, Swathi Bhat, and Sudarshan Karigoudar. "ResNet-18 based Multi-Task Visual Inference and Adaptive Control for an Edge-Deployed Autonomous Robot." Frontiers in Robotics and AI 12 (2025): 1680285.