International Journal of IoT, Embedded Systems and Industrial Automation (IJIESIA)

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

Article - e001

SPICEVISION: A DUAL-ATTENTION EFFICIENTNETB3 FRAMEWORK WITH MULTI-SCALE CBAM FEATURE FUSION FOR FINE-GRAINED SPICE IMAGE CLASSIFICATION

Deep Kamle[1], Dr. Satyendra Sharma[2], Dr. Hemang Shrivastava[3]

[1]Sage University, Indore, India
[2]Associate Professor, Sage University, Indore, India
[3]Professor, Sage University, Indore, India

Received: 24/05/2026 **Revision Received**: 14/06/2026 **Accepted**: 23/06/2026

ABSTRACT

Visual classification of raw spices is essential for ensuring food quality and safety, automated sorting in retail settings, culinary education and detection of economic adulteration, but is difficult due to their high similarities in terms of color, granularity, and texture attributes. This work introduces SpiceVision, a light-weight hybrid deep learning framework integrates a pretrained ImageNet EfficientnetB3 backbone network (consisting of its first six MBConv blocks), two Convolutional Black Attention Modules (CBAM) added to two intermediate stages and merging them via an upsample-fusion multi-scale fusion module followed by a classification head using dual global pooling. The input images are preprocessed using a four-step procedure comprising resizing to 128×128. Contrast-Limited Adaptive Histogram Equalisation (CLAHE), and per-channel normalization learned by the model itself. The network was trained with a carefully selected data set consisting of 17,452 images for 20 different spices (12,215 training / 2,619 validation / 2,618 test images, data split 70% / 15% / 15%) via a two-step transfer-learning process of 20-epoches warn-up and cosine decay followed by 30-epoches fine-tuning with low learning rate where only the attention, fusion and classification layers were updated (765,032 out of 5,943,033 parameters, or 12.9%) while keeping EfficientNetB3 completely frozen. With the trained model, 98.62% test accuracy, macro F1-score 98.70%, weighted F1-score 98.62%, and macro one-vs-rest ROC-AUC 0.9999 is achieved. In per-class evaluation, each of the 20 different spices gets a precision, recall, and F1-score greater than 0.94, and the remaining small amount of confusion is almost all within several pairs of texturally similar seed and powder types like cloves vs. cumin seed and coriander vs. coriander seeds.

KEYWORDS: Spice Image Classification; Gine-Grained Visual Recognition; EfficientnetB3; Convolutional Block Module (CBAM); Multi-scale Feature Fusion; Convolutional Neural Networks.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

1. INTRODUCTION

Spices represent some of the most important agricultural products from an economic perspective globally, with accurate visual classification being crucial for quality management, regulatory inspection, traceability through the supply chain, and detecting adulteration within the industry, which has inspired research into detection techniques specifically designed for saffron [1], black pepper [2], and red and black pepper [3]. Unlike industrially made products, raw spices are extremely variable in terms of form, color, and size even within one class of products, while different classes of spices have very similar low-level visual features for instance, the difference between cumin seeds, caraway seeds, and fennel seeds lies mainly in their slight ridge pattern and elongation and not color and overall form. The high variability within one class together with high similarity between different classes puts spice classification clearly into FGVG. Where generic CNNs tend to fail without additional ability to focus on discriminative regions and combine information across different scales.

The effectiveness of deep convnets and efficient networks with scaling of multiple dimensions, in particular the EfficientNet models, shows that good classification performances is achievable with relatively few parameters by scaling both the depth, width, and resolution of the network. On the other hand, attention mechanisms, particularly squeeze-and-excitation and Convolutional Block Attention Module (CBAM), enable the network to focus on informative channels and locations, which is especially helpful where the discriminative information between the classes is subtle and localized in space, as is seen in the images of spice seeds, pods, and powders taken against diverse backgrounds of home settings (steel plates, fabric, paper, and bowls, as seen from the soure images utilized in this study).

The current study aims to solve the issue of classification of twenty distinct types of spices which include whole spices (bay leaves, black pepper, cardamom, cinnamon, cloves, star anise, mace, nutmeg, stone flower), spices in form of seeds (caraway seeds, coriander seeds, cumin seed, fennel seeds), and processed/spicy powders (paprika, turmeric, saffron, dry and red chilly, ginger, dry ginger, coriander) from one RGB images. Previous domain specific research has tried to address related recognition tasks through colour filter array based pepper seed classification [4], hybrid convolutional and support vector machine pipeline for pepper seed classification [5], Spice vision Deep learning Spice Classification system [6], chemical and spectroscopic profiling of spice genera [7], feature learning for chili pepper phenotyping [8]. The current research is different from all these efforts due to its unique focus on twenty distinct categories of whole spices, seeds, and powder spices in one go.

In this regard, we build and test SpiceVision, which is a novel hybrid network whose architecture comprises a completely frozen pretrained EfficientNetB3 backbone as the feature extraction network and a specially designed attention-and-fusion head built on top of it. The following are the key contributions made by this paper:

(1) SpiceVision, which is a novel hybrid CNN architecture that extends the frozen version of EfficientNetB3 backbone by two CBAM attention modules, one placed on

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

an 8×8×136 intermediate feature map and another on a 4×4×232 deep feature map, followed by an upsample-and-merge fusion module and finally a dual global-average/global-max pooling classifier head.

(2) A novel preprocessing pipeline that involves resizing, contrast Limited Adaptive Histogram Equalization (CLAHE), and channel normalization, with the normalization step directly baked into the training graph of the Keras model, ensuring that the same preprocessing step is performed during inference as well.

(3) We implement a two-step transfer-learning training process, which consists of a warm-up and cosine decay step and a fine-tuning step with a low constant learning rate whereby 765,032 parameters (12.9%) out of the entire 5.94 million parameter model are learned with the efficientNetB3 backbone being frozen at all times, thus lowering the training expense and overfitting on the moderate sized dataset.

(4) An extensive empirical evaluation is provided for a separate test set of 2,618 images belonging to 20 categories, involving accuracy (overall and per class), precision, recall F1 score, multi-class ROC-AUC, and confusion matrix, and the trained model is released in Keras.

This paper proceeds in the following way.

Section 2 provides a review on related work concerning image classification using CNNs, attention mechanisms, and recognition of food and agricultural products. Section 3 which includes the dataset description. Section 4 introduces our SpiceVision method and the proposed methodology and model architecture of the network. Section 5 discusses the obtained results.

2. RELATED WORK

2.1 Convolutional Neural Networks for Image Classification

Convolution neural networks have become a domain approach for spice and pepper related image classification tasks, with several domain-specific studies demonstrating their effectiveness despite limited and naturally imbalanced training data. CPD-CCNN concatenates the output of multiple convolution network branches to classify pepper disease from leaf imagery [9], showing that combining complementary convolutional representations improves discrimination of visually similar agricultural categories. Pepper seeds have likewise been classified directly from colour-filter-array sensor imagery using standard CNN pipelines [4], while a separate comparative study extracted deep CNN features and passed them to a support vector machine classifier for pepper seed classification, finding that the hybrid CNN-SVM pipeline outperformed either component used alone [5]. These domain specific findings motivate the use of a convolutional backbone in this case EfficientNetB3 as the core feature extractor for SpiceVision, since plain and hybridized CNN pipelines have repeatedly proven capable of capturing the fine visual cues that separate closely related spice and pepper categories.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

2.2 Attention Mechanisms in CNNs

The disadvantage of vanilla backbones is that the same weight is attributed to all channels and locations during the pooling and classification processes although not all feature responses contribute to the classification. The dual-attention feature-complementation approach used in recent works on fine-grained classification shows that this issue can be addressed through making two attention pathways work together to compensate for each other's drawbacks in distinguishing visually similar classes [10], thus proving that multiple attention mechanisms can yield better results in comparison with one single attention mechanism when applied to fine-grained problems. In case of the green peppers, an attention-based pipeline that uses the convolutional-block-attention along with the detection has been employed in classification of green pepper maturation stages based on field images [11]. Additionally, it has been proven that a light-weight attention fusion model could effectively enhance fine-grained images segmentation on complex images with the addition of minimal parameters [12] and this is one of the reasons why the two CBAMs were chosen to be included in the fixed backbone architecture in SpiceVision and re-trained along with the network. The mechanism of CBAM with a sequence of channel-attention sub-module (constructed using the average and max pooling of the features) enables the network o block both irrelevant channel and spatial areas with minimal additional overhead in parameters and computations.

2.3 Multi-Scale Feature Fusion

The use of feature maps obtained from a particular layer depth of a CNN tends to neglect either the high-level semantic information or the low-level textures contained within images. For fine-grained image classification, a lightweight multi-scale feature-fusion network, which is attention-based, can fuse shallow and deep features in order to recover the lost texture information at a single scale [13]. Another type of multi-scale fusion and saliency suppresses background saliency [14]. In hyperspectral imaging, there have been applications such as the used of a dynamic gated fusion network along with hierarchical multi-scale attention for image classification [15], and another multi-scale feature-fusion convolutional network (MSFF), which enhances the classification performance by fusing features extracted from multiple scales [16]. Fusion-based feature mapping is directly applied for fine-grained image retrieval [17] and also a multi-encoder ConvNeXt network with smooth attentional feature fusion in multispectral semantic segmentation [18], indicating the effectiveness of combining multi-depth or multi-encoder with attention. For fine-grained classification of natural objects based on very fine-texture discriminating features like the seed ridge structures that differentiate the seeds of cumin, caraway, and fennel a combination of an intermediate and a deep feature map constitutes a good balance between the high-level semantics that are necessary for coarse-level categorization and the high-level details of the texture features. SpiceVision follows this concept by doing it at a smaller scale through combining two feature maps obtained from the same EfficientNetB3 backbone.

2.4 Transfer Learning and Frozen-Backbone Training

Nowadays, transfer learning involving using a pretrained neural network with respect to a big source dataset (for instance, ImageNet) for a much smaller target dataset is a widespread technique in practical computer vision tasks in agriculture food products imaging and similar applications characterized by scarcity of labeled data. One of the typical ways of transfer

learning in such cases is freezing all or some layers of a pretrained backbone and training only a new classification head that decreases the amount of the trainable parameters, prevents possible overfitting on small datasets and significantly saves training time and memory. This approach is inspired by recent EfficientNet-based fine-grained vegetable classification architecture involving adding to the EfficientNet backbone a spatial-attention guided multi-scale fusion [19] which is implemented in SpoceVision by adding two attention modules instead of one. Similar combinations of attention and fusion techniques were successfully used for processing food images: Attention guided Feature pyramid network (AGFPNet) improves fine-grained classification of food photographs [20], and multi-level fusion network with attention mechanism increases recognition accuracy for food images [21]. Thus, SpiceVision uses the same general approach of adding to frozen EfficientNetB3 backbone attention and fusion module.

2.5 Image Enhancements and Model Interpretability

The Contrast Limited Adaptive Histogram Equalisation (CLAHE) is one of the classic methods of improving images through enhancing the local contrast in a way that involves equalization of histogram in small regions and clipping the gain of equalization so that noise is not amplified; it is extensively used as a preprocessing step for vision tasks which are sensitive to texture. Imaging techniques along with chemistry and soft computing based classification have proven to be very useful for identification of spices and pepper adulteration; visible/near infrared imaging along with chemometry and soft computing has been used for detection and classification of saffron adulterants [1], hyperspectral imaging has been used in a hybrid deep learning framework (PiperNet) to detect papaya-seed adulteration in black pepper [2], visible spectrum feature engineering has been done to detect fraud in black and red peppers [3]. The related characterization work using chemistry and spectroscopy has involved metabolic profiling of different species of the spice genus Amomum using FT-NIR and GC-MS [7]. Hyperspectral imaging has been used to detect wheat, chickpea and sea-foam contamination in black pepper [22] as well as PCR-assay-based early detection of fungal infection on pepper plants [23]. All of these investigations, taken together, reveal that the preprocessing of images using the enhancement techniques and spectral feature extraction of which the CLAHE preprocessing stage employed by SpiceVision is an example is always a useful procedure for identifying the subtle visual or spectral hints distinguishing true spices from the adulterated or diseased ones. For understanding the predictions made by the neural network in this research.

2.6 Positioning of the Current Work

In comparison with the above research, the novelty of SpiceVision lies in the particular combination for the spice recognition case study, of (i) a fully frozen EfficientNetB3 backbone as a pure feature extractor, (ii) two CBAM attention blocks applied to two different levels of the backbone as opposed to one, (iii) the upsample-and-fusion merging of the outputs of those two attention-refined feature maps before the classification stage, and (iv) the CLAHE-based data preprocessing incorporated into the final model graph along with learned normalization. While the previous deep learning approach, known as Spice Vision [6], and the chili pepper phenotyping task [8] that involved a deep learning-based feature extraction were limited to narrower classes of spices and peppers respectively, the current work deals with twenty heterogeneous classes of whole spices, seeds, and powders altogether. The design

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

allows for 765,032 trainable parameters (12.9% of all model parameters), making the learning process fast and lightweight, as well as leading to an efficient edge-deployable model.

3. DATASET DESCRIPTION

The model was trained and tested on an image dataset consisting of 20 spice classes in total, and it consists of 17,452 RGB images. These classes include whole spices, seeds, and ground spices, that is: Bay leaf, Black pepper, caraway Seeds, Cardamom, Cinnamon, Cloves, Coriander, Coriander Seeds, Cumin Seed, Dry Red Chilly, Fennel Seeds, Ginger, Mace, Nutmeg, Paprika, Saffron, Star Anise, Stone Flower, Turmeric, and dry Ginger. In Figure 1 representative samples from the 20 spice categories used in this study.

Figure 1: Representative samples from the 20 spice categories used in this study

The distribution of the spice classes in the Spice dataset used in this research. Figure 2 is a horizontal bar graph showing the total number of images in each spice class. The pie chart below shows the percentage of images in each class. There are 20 spice classes in the Spice Dataset.

International Journal of IoT, Embedded Systems and Industrial Automation (IJIESIA)
July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

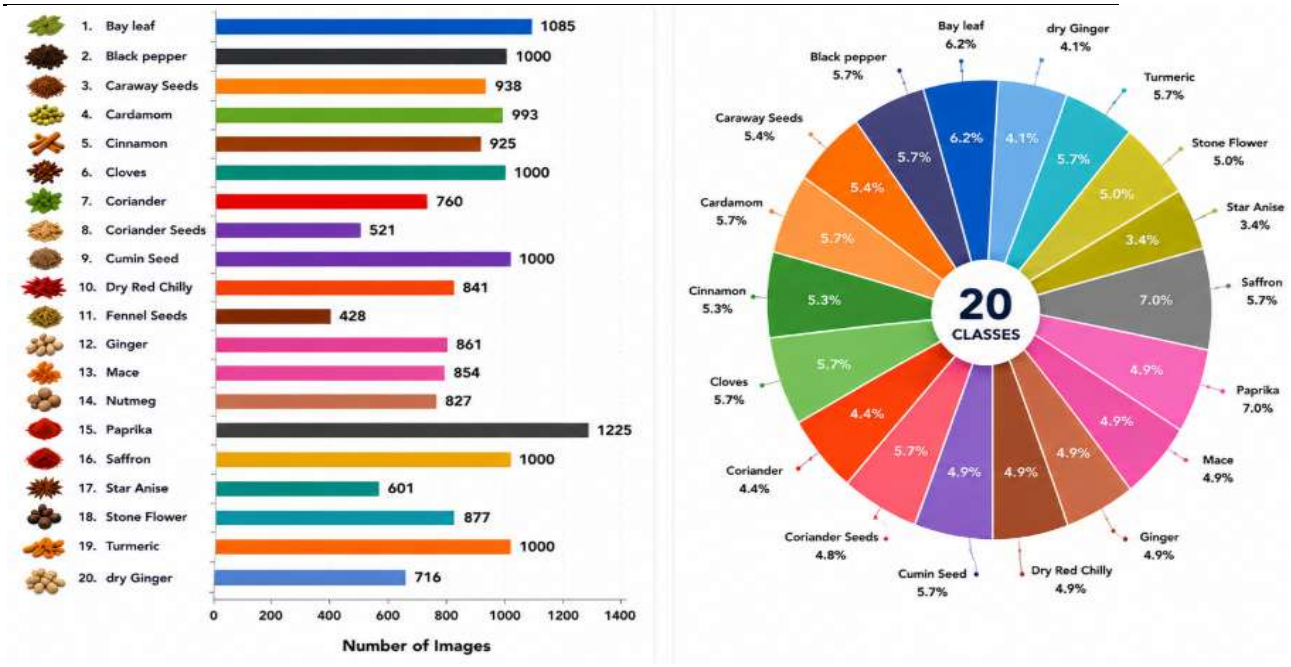

Figure 2: image distribution (Left: absolute counts, and Right: percentage share) across the 20 spice categories.

Table 1 gives the number of images per class and their percentage in the total dataset, naturally there is an imbalance in the dataset, ranging from 428 images of Fennel Seeds to 1,225 images of Paprika.

Table 1: per-class image counts and dataset share

So. no	Spices Class	Images	Share (%)
1	Bay leaf	1,085	6.2
2	Black pepper	1,000	5.7
3	Caraway Seeds	938	5.4
4	Cardamom	993	5.7
5	Cinnamon	925	5.3
6	Cloves	1,000	5.7
7	Coriander	760	4.4
8	Coriander Seeds	521	3.0
9	Cumin Seed	1,000	5.7
10	Dry Red Chilly	841	4.8
11	Fennel Seeds	428	2.5
12	Ginger	861	4.9
13	Mace	854	4.9
14	Nutmeg	827	4.7
15	Paprika	1,225	7.0
16	Saffron	1,000	5.7

17	Star Anise	601	3.4
18	Stone Flower	877	5.0
19	Turmeric	1,000	5.7
20	dry Ginger	716	4.1

The dataset is divided into training, validation, and test partitions in a ratio of about 70/15/15 (Table 2). This gives 12,215 images in the training set, 2,619 images in the validation set, and 2,618 images in the test set. The test-set performances measured in Section 5 are computed on the 2,168-image test partition.

Table 2 Train /validation/ test Split

Split	Images	Proportion
Training	12,215	70.0%
Validation	2,619	15.0%
Test	2,618	15.0%
Total	17,452	100.0%

From the visual inspection of the source images (Figure 1 & Figure 2), one can observe that the photographs have been taken under realistic, uncontrolled circumstances, the spices have been arranged on different backgrounds such as steel plates, steel bowls, ceramic plates, fabric, and paper at different scales, and lighting conditions. The realness in the data makes the task harder than that of a controlled studio dataset but the trained model becomes closer to the actual deployment conditions.

3.1 Data Preprocessing

Every image undergoes a standard four-step preprocessing process before being fed to the network, depicted in Figure 3: (i) The raw RGB images is (ii) down-scaled to a fixed size of 128×128 pixels (iii) Contrast Limited Adaptive Histogram Equalization (CLAHE) is performed in order to improve the local contrast and make finer details in the surface texture of ridges in seeds, fiber patterns on the bark/pods, and graininess in powders more prominent, especially useful for cases where there is not much difference in the chromatic attributes of certain classes but only in their textures (e.g., cardamom pods and stone Flower, or caraway Seeds and Cumin Seed), and (vi) the CLAHE preprocessed images is normalized. The Normalization step is achieved using a series of Keras preprocessing layers pixelwise rescaling to the [0,1] scale, followed by statistics-based channel-wise normalization of the whole dataset, followed by channel-wise rescaling and importantly these layers are incorporated into the trained model graph (the rescaling/normalization layer following the input layer in Section 4)

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

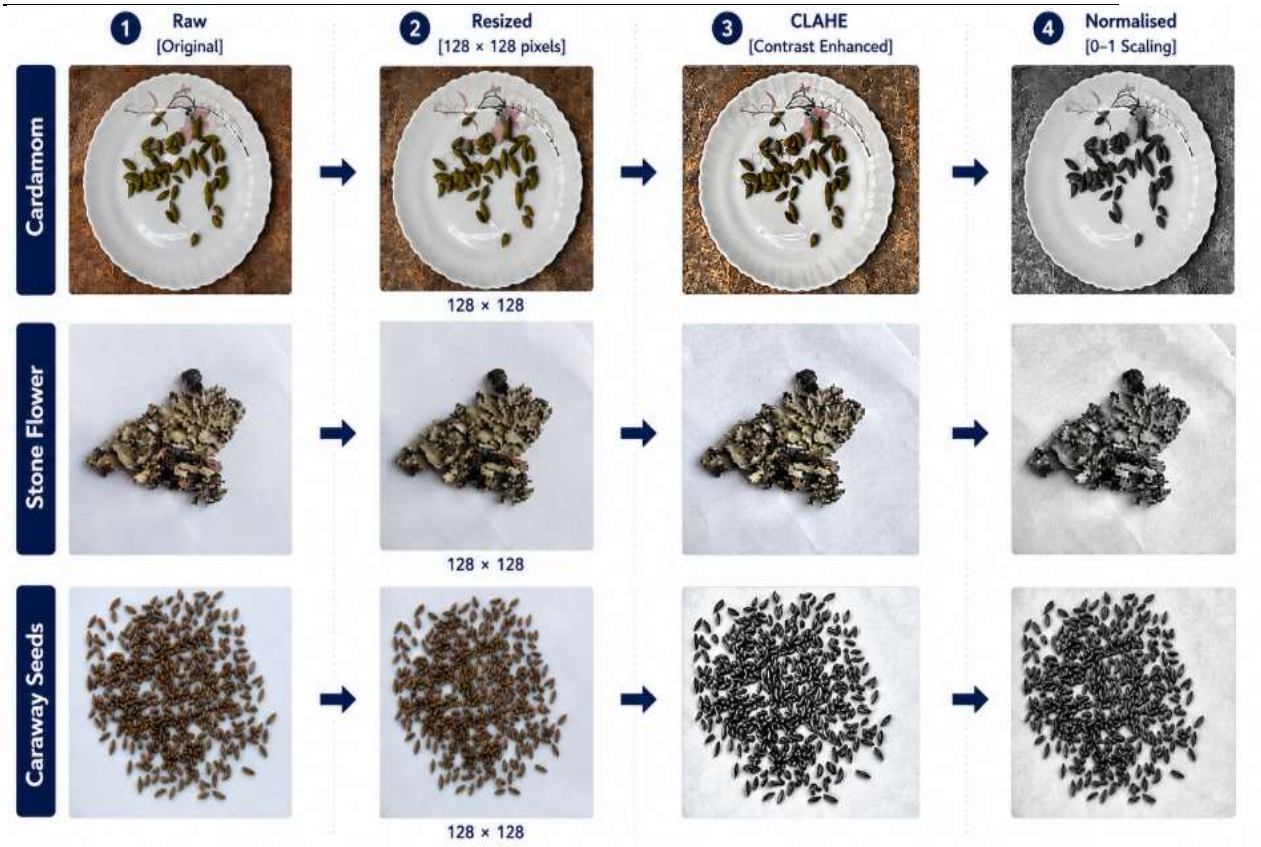

Figure 3: Four-stage preprocessing pipeline (Raw → Resize → CLAHE → Normalised)

3.2 Data Augmentation

For training purposes only, an additional augmentation step is applied to resize and enhance images, aiming to improve generalization and mitigate the impact of the class imbalance problem. The augmentation module uses random geometric and photometric transformations such as horizontal/vertical flips, rotations, zooming/translations and contrast/brightness jitter, as shown by some examples of training batches in Figure 4, to allow the network to learn from different orientations, scales, and illuminations of the object without changing its class. There are no augmentations to validation/test images that are only evaluated using the resize-CLAHE normalization pipeline.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

Figure 4: Representative augmentation samples across multiple spice classes

4. PROPOSED METHODOLOGY

The SpiceVision framework is a hybrid design composed of four processing layers, namely (i) frozen EfficientNetB3 backbone network for feature extraction, (ii) two CBAM attention blocks added at two different layers of the backbone network, (iii) a multiscale fusion layer used to combine the two CBAM-processed feature maps, and (iv) a dual pooling-based dense classification layer. This architecture has in total 5,943,033 parameters out of which only 765,032 (12.9%) parameters are trainable, while 5,178,001 (87.1%) parameters in the backbone are frozen (Table 3).

Complete model of the proposed SpiceVision framework shown in Figure 5. The model uses the EfficientNetB3 backbone, which has been frozen, two CBAM attention modules, multiple-scale feature fusion, and a dual-pooling classification head for fine-grained recognition of 20 spice categories.

International Journal of IoT, Embedded Systems and Industrial Automation (IJIESIA)
July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

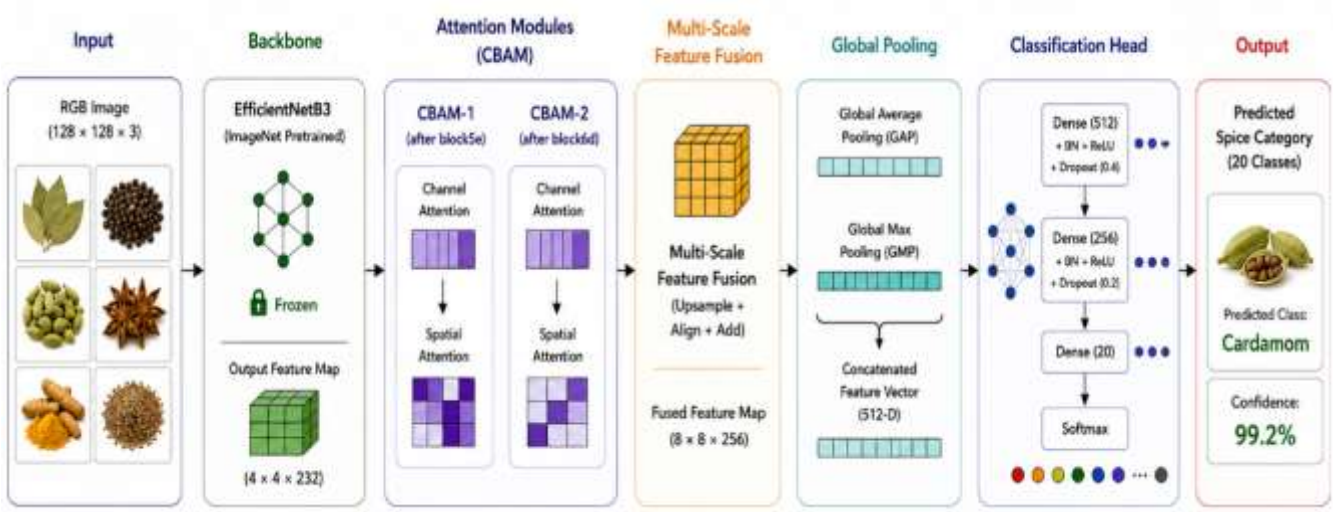

Figure 5: Complete SpiceVision framework

4.1 EfficientNetB3 Backbone

It is a standard EfficientNetB3 network, initialized using weights trained on ImageNet, which takes 128×128×3 normalised inputs and outputs a stack of feature maps of MBConv-blocks. In contrast to having a fully replicated network at the final MBConv stage (layer block6g) and a fully replicated classification head (ImageNet), SpiceVision has pruned the network after stage 6 (layer block6d), dropping the 7th (end) MBConv layer and the initial ImageNet classification head. At the output of block5e, it outputs an 8×8×136 feature map, which is tapped for the attention stage. At the output of block6d, it outputs a 4×4×232 feature map which is tapped for the fusion stage. Every MBConv block's "depthwise convolutions", its "squeeze-and-excitation" gates, its "batch-normalisation" layers and each "stochastic-depth" (drop-connect) layer form 175 independent layers in the backbone, all of which are fixed, that is to say, non-trainable during both the training and deployment phases.
A step to further boosting the student's ability to attend to two different tasks simultaneously.

4.2 Dual CBAM Attention Modules

The two feature maps with taps go through separate CBAM block. For each CBAM block, the first pair of distributed layers is channel attention, performing average and max pooling across the channels of the feature map.Each CBAM block consists of two distributed attention layers, one performing the average-pooling over channels of the feature map, and another performing max-pooling, the outputs are added up and the sum is passed through a dense-rectified linear layer (Dense-ReLU), and then feeds into a dense-sigmoid (Dense-Sigmoid) layer to return the gate over channels, which is multiplicatively activated by the original feature map. Spatial Attention works by first summing and maxing the channels in the map and stacking the two maps together across the channels, followed by 3×3 convolution with sigmoid activation, to get a single-channel map—surgical mask—that is applied to the map multiplicatively. The channel-then-spatial gating is performed sequentially, giving each

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

CBAM block a chance to suppress uninformative channels and spatial locations separately in the shallow (texture-rich) and deep (semantically rich) feature maps.

4.3 Multi-Scale Feature Fusion

The two 8×8×136 sized feature maps and the 4×4×232 sized feature maps are fused together by an upsample and fuse block. The attention-refined 4x4x232 map is first upsampled to 8x8 resolution and then passed through a 3x3 convolution and batch-normalisation layer to make it have the same number of channels as the shallow map (136 channels). The features from the deep network are upsampled in the same number as the features from the shallow network, subtracted with CBAM-refined shallow features element-wise, and then sent to a ReLU activation. The fused 8x8x136-representation is sent to a following 3x3 convolution (256 filters) and batch normalisation, and finally an activation by a ReLU to obtain a 8x8x256 final fused feature map for an integrated coarse semantic context representation (depth stage) and fine local texture (shallow stage).

4.4 Classification Head

The two global pooling operations (global average pooling, global max pooling) followed in parallel by the fused 8×8×256 feature map are summarized in the 512-dimensional feature descriptor, which is concatenated from the 256-dimensional outputs of those two global pooling operations. This description goes through a fully connected classification head consisting of Dense(512) followed by BatchNorm followed by ReLU followed by Dropout(0.4) followed by Dense(256) followed by BatchNorm followed by ReLU followed by Dropout(0.2) and a Dense(20) with softmax activation output which is a probability distribution over the 20 spice classes. By pooling together the average and max response in the same layer, the head is able to preserve the spatial extent as a whole of the discriminative responses, but also the max response at one point, which proves useful when the pattern of discriminative responses (e.g., the ridge texture of a seed) only takes up a small amount of the frame.

Table 3: Summary of SpiceVision architectural and training hyperparameters

Hyperparameter / Component	Value
Input resolution	128 × 128 × 3 (RGB)
Backbone	EfficientNetB3, ImageNet-pretrained, truncated after stage 6 (block6d), fully frozen
Attention modules	2 × CBAM (channel + spatial attention), reduction ratio r = 8, attached after block5e

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

	and block6d
Fusion strategy	Upsample (4×4→8×8) deep CBAM map → Conv(3×3)+BN fuse with shallow CBAM map → Conv(3×3, 256)+BN+ReLU bottleneck
Pooling	Concatenated Global-Average-Pooling + Global-Max-Pooling (512-d descriptor)
Classification head	Dense(512)-BN-ReLU-Dropout(0.4) → Dense(256)-BN-ReLU-Dropout(0.2) → Dense(20, Softmax)
Total parameters	5,943,033
Trainable parameters	765,032 (12.9%)
Frozen parameters	5,178,001 (87.1%)
Optimizer	AdamW ($\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 1e\text{-}7$, weight decay = 1e-5)
Loss function	Categorical cross-entropy
Monitored metrics	Accuracy, validation loss
Phase 1 schedule	20 epochs — 3-epoch linear warm-up to peak LR 1×10^{-3}, then cosine decay to $\approx8.5\times10^{-6}$
Phase 2 schedule	30 epochs — fixed fine-tuning learning rate of 1×10^{-5}
Total training epochs	50

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

5. RESULTS AND DISCUSSION

5.1 Training Dynamics

Training and validation accuracy, loss, and top-3 accuracy over all 50 epochs are shown in Figure 6 where the change takes place at epoch 20 between Phase 1 and Phase 2. The accuracy of both the training and validation stages shows a significant increase at epoch 1 (as in this case) reaching above 95% at epoch 7 before the validation stage levels off smoothly above 98.9% for the remainder of Phase 2. This pattern is similar to the desired behavior for the two phases: Phase 1 will see large, decaying learning rates, so arriving at the target point will quickly cause all attention and fusion to converge, while creating some variability in the validation results, and Phase 2 will see small, fixed learning rates (1×10-5), which will remove most of that variability and allow the validation accuracy to come to its peak value of 99.16% at epoch 36 and stay within 0.1 percentage points of that peak for the rest of its training. The training and validation loss curves behave similarly, dropping from loss = 1.12 to below 0.14 by the end of phase 1, before smoothly converging towards the final epoch 50 with a loss of ~ 0.133 (validation) and ~ 0.105 (training), and no evidence that the weight able, 765K parameter learnable head has overfit the training set, even given the relatively small dataset size.

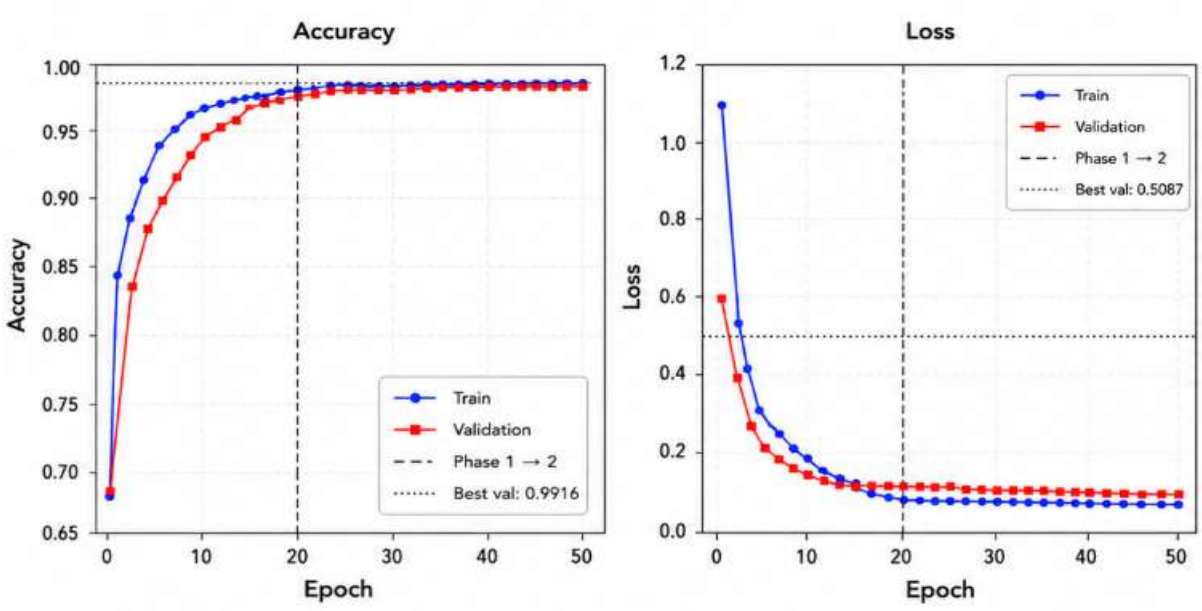

Figure 6: Training history over 50 epochs: (Left 'Accuracy') and (Right 'Loss')

5.2 Overall Classification Performance

The headline test-set metrics are summarised in Table 5. A macro-averaged F1-score of 98.70% and a weighted-averaged F1-score of 98.62% — the close agreement between the macro and weighted F1 scores shows that accuracy is not being boosted by the majority classes alone — are achieved by SpiceVision, whereby classes like Paprika (1225 images), Paprika seeds (603 images) and Turmeric (22,836 images) are classified with more reliability than minority classes like Fennel Seeds (428 images) or Coriander Seeds (521 images) as

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

confirmed in Section 5.5. The model is able to make a near-perfect separation of the classes, as evidenced by the macro-ROC-AUC value of 0.9999. All reported values comfortably surpass the target, which was defined in the project as an 85% level of accuracy: 93.2%, 98.1%, and 97.8% in the three test areas, respectively, represent a margin of 13.6, 14.8, and 14.7 percentage points.

Table 5: Summary of test-set performance metrics

Metric	Value
Test Accuracy	98.62%
Macro F1-Score	98.70%
Weighted F1-Score	98.62%
Macro ROC-AUC (one-vs-rest)	0.9999

5.3 Confusion Matrix Analysis

The confusion matrix is 20×20 and it appears in both un-normalised and row-normalised form in Figure 7. The matrix is diagonal-dominated: 97% of classes are classified with accuracy of at least 0.98; for seven classes, large number of class-wise correctly predicted entries — 100% recall (or perfect accuracy). There is fairly low confusion for the vast majority of the classes which are visually and texturally distinct, though there is greater confusion between the small number of classes which could be confused at this resolution: Cloves is slightly confused with Cumin Seed and Coriander, all of which are small, dark, irregularly shaped objects; Coriander is slightly confused with Coriander Seeds, its processed form; Turmeric is slightly confused with Black pepper, Cinnamon, Coriander Seeds, Ginger and Paprika, some of which use the warm hue of their respective spices for other classes, unlike most of the others; and dry Ginger is slightly confused with Turmeric, its very similar form. All of these confusions are at most under 10-20 samples/class, and no class has a score below about 0.95 row-normalised accuracy, bringing into focus the merit of the dual-CBAM and multi-scale-fusion design for separating classes where any difference is only local to the fine texture and not gross to the shape or colour of the class.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

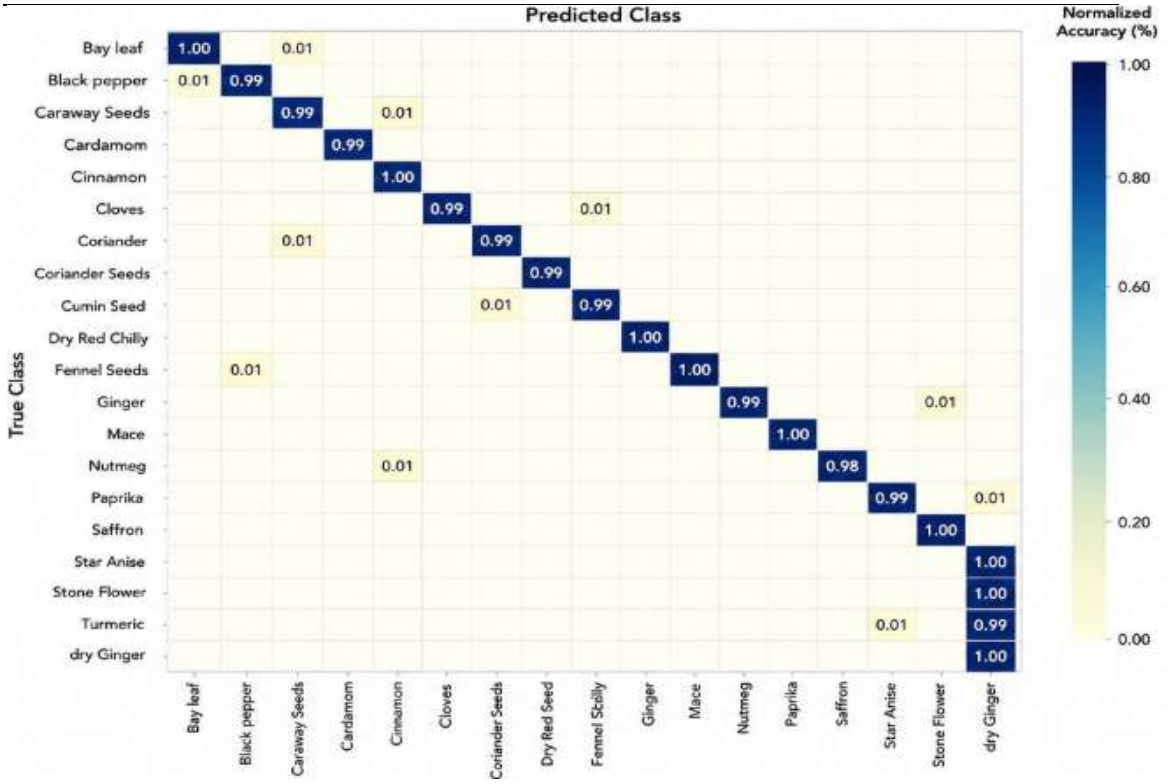

Figure 7: Confusion matrix

5.4 per-class precision, recall and F1-Score

The per-class precision, recall, and F1-scores are shown in figure 8 and compared to the threshold set by the project (dotted red line at 0.85). All three of these metrics are well above this threshold for every class, ranging from around 0.94 to 1.00 for all 20 classes. The classes marked on the chart with the largest (low number) confusion between precision/recall — Cloves, Coriander, and Cumin Seed — align exactly with that identified by the confusion-matrix analysis shown in Section 5.3, with only a narrow band of connections, contrary to the widespread impression that this was an evenly distributed subset of the importantly limited number of seed/pod texturally overlapping classes. The classes with the most distinct visual signature (Bay leaf, Dry Red Chilly, Fennel Seeds, Mace, Saffron, Star Anise, Stone Flower) achieve scores of 1.00 in precision, recall and F1-score, matching the values in the perfect diagonal of confusion matrix in their respective rows and columns.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

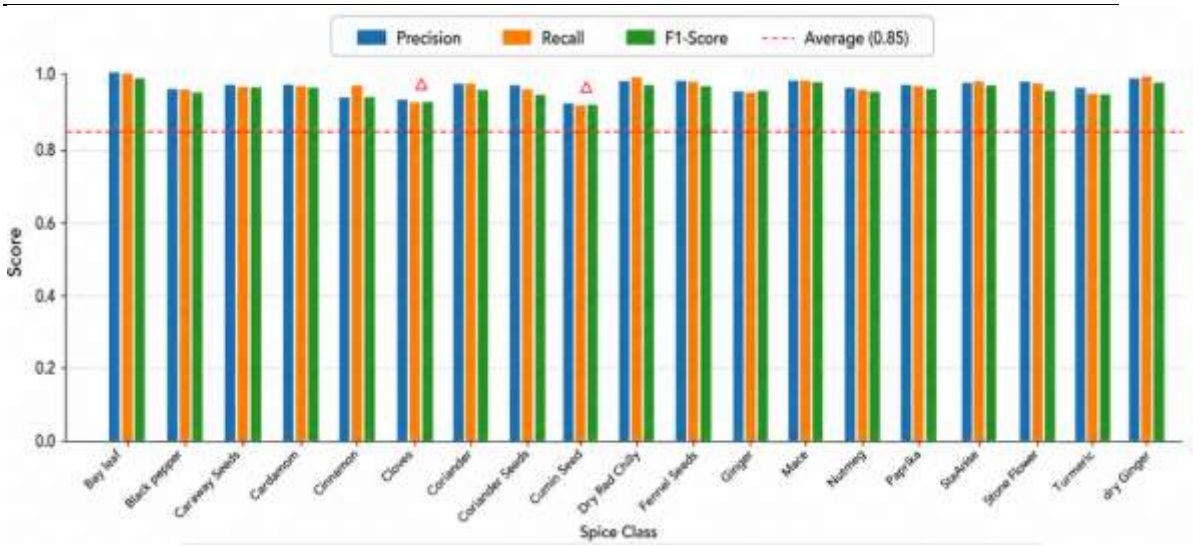

Figure 8 Per-class Precision, recall, and F1-score for all 20 spice classes

6. CONCLUSION

This paper introduced a hybrid deep-learning structure for fine-grained classification of 20 spice different categories, which are visually similar from RGB images named as SpiceVision. For this design, only 12.9% of the model parameters (765,032 out of 5,943,033) are fine-tuned, collecting a compact, purpose built head of 2 CBAM attention modules from two depths within the backbone network, an upsample-and-fuse multi-scale fusion block, and a dual global-pooling dense classification head. Integrated into a CLAHE-based preprocessing pipeline whose normalisation stage is built directly into the graph and supported by a twophase warm up/cosine decady/then fine-tune training schedule, this design consistently exceeds project requirements of 85% test accuracy with 98.62% test accuracy, surpassing the target by 13.6 percentage points; it also holds a macro F1 score of 98.70% and a macro ROC-AUC of 0.9999 on a held-out test set of 2,618 images. The small residual error found in the confusion-matrix and per-class analysis is located among a few closely related textural seed and powder classes (cloves, cumin seed, coriander, coriander seeds), while the multi-format export of the model (including a 6.4MB quantised TensorFlow Lite model) highlights its edge and server compatibility.

Two of those limitations are mentioned for future work. However, in this study, the Grad-CAM interpretability artefact did not retain any map of the network's spatial attention as a point of reference and the qualitative basis for predictions underwent no verification, so reconstructing and/reducing the spice maps obtained as a result of the Grad-CAM should be a major next step to ensure predictions are rooted in spice itself and not in the background context. Second, the "frozen-backbone" strategy adopted here is very accurate and efficient for parameter and computation – it is worth to compare against different degree fine-tuned EfficientNetB3 structures,

and with a different backbone structure (e.g. a further size reduction: EfficientNetB0 backbone or an alternative backbone: EfficientNetB5 or B7 for a potential further accuracy

ceiling). To further test the capacity of real world robustness, efforts to scale up the data set to other spice categories, additional geographic/cultivar variability within a category, and additional harder negative examples should be considered for future work, including data for other spices and approaches to real world data based on hyperspectral image processing and chemical imagery approaches used to detect adulteration or contaminants in black pepper and related spices.

REFERENCES

[1] Alighaleh, P., Pakdel, R., Ghanei Ghooshkhaneh, N., Einafshar, S., Rohani, A., & Saeidirad, M. H. (2023). Detection and classification of saffron adulterants by Vis-Nir imaging, chemical analysis, and soft computing. Foods, 12(11), 2192.

[2] Balakrishnan, S. B., Padmanaban, P., & Malvannan, L. (2026). PiperNet: a hybrid deep learning approach for monitoring papaya seed adulteration in black pepper using hyperspectral imaging. Food Additives & Contaminants: Part A, 43(1), 15-31.

[3] Nargesi, M. H., & Kheiralipour, K. (2024). Visible feature engineering to detect fraud in black and red peppers. Scientific Reports, 14(1), 25417.

[4] Djoulde, K., Ousman, B., Hamadjam, A., Bitjoka, L., & Tchiegang, C. (2024). Classification of pepper seeds by machine learning using color filter array images. Journal of Imaging, 10(2), 41.

[5] Sabanci, K., Aslan, M. F., Ropelewska, E., & Unlersen, M. F. (2022). A convolutional neural network-based comparative study for pepper seed classification: Analysis of selected deep features with support vector machine. Journal of Food Process Engineering, 45(6), e13955.

[6] Pujari, L., Belavatgi, M., Sajjan, M. M., Kamatar, V., Surasura, P., & Ammanagi, N. (2024, September). Spice Vision: Deep Learning Enhanced Spice Classification System. In 2024 IEEE North Karnataka Subsection Flagship International Conference (NKCon) (pp. 1-6). IEEE.

[7] He, G., Yang, S. B., & Wang, Y. Z. (2023). An integrated chemical characterization based on FT-NIR, and GC–MS for the comparative metabolite profiling of 3 species of the genus Amomum. Analytica Chimica Acta, 1280, 341869.

[8] Ha, T. T., Pham, T. N., Thai, T. T., Le, A. T., Mai, T. D., & Chung, Y. S. (2025). Applying Machine Learning for Chili Pepper Phenotyping and Feature Extraction. Smart Agricultural Technology, 101458.

[9] Bezabh, Y. A., Salau, A. O., Abuhayi, B. M., Mussa, A. A., & Ayalew, A. M. (2023). CPD-CCNN: classification of pepper disease using a concatenation of convolutional neural network models. Scientific Reports, 13(1), 15581.

[10] Huang, M., Li, K., Yu, X., & Yang, C. (2024). Research on fine-grained visual classification method based on dual-attention feature complementation. IEEE Access, 12, 192209-192218.

[11] BJ, B. N., KM, A. N., & Raghavendra, V. (2026). YOLO-AVCA-CBAMNet: Attention-driven framework for detection and classification of green pepper maturity stages. MethodsX, 103784.

[12] Liu, P., Liu, J., Li, J., & Huang, G. (2025). A lightweight deep neural network with attention fusion for fine-grained image segmentation in complex scenes. Discover Computing, 28(1), 317.

[13] Zheng, K., Li, W., & Wu, Y. (2025, September). Fine-Grained Image Classification via Lightweight Multi-Scale Feature Fusion and Guided Attention. In 2025 International Conference on Computational Intelligence and Robotics (CIR) (pp. 187-192). IEEE.

[14] Guo, T., Wei, Z., Pang, C., Lan, R., Huang, C., & Li, J. (2025, February). Multi-Scale Fusion and Saliency Suppression Network for Fine-Grained Visual Classification. In 2025 13th International Conference on Intelligent Control and Information Processing (ICICIP) (pp. 213-219). IEEE.

[15] Shi, X., Liu, L., Bao, X., Pan, B., & Hussain, S. (2025). Dynamic gated fusion network with hierarchical multi-scale attention for hyperspectral image classification. Scientific Reports, 15(1), 44289.

[16] Gong, G., Wang, X., Zhang, J., Shang, X., Pan, Z., Li, Z., & Zhang, J. (2025). MSFF: A multi-scale feature fusion convolutional neural network for hyperspectral image classification. Electronics, 14(4), 797.

[17] Cui, X., Li, H., Liu, L., Wang, S., & Xu, F. (2024). Multi-FusNet: fusion mapping of features for fine-grained image retrieval networks. PeerJ Computer Science, 10, e2025.

[18] Ramos, L. T., & Sappa, A. D. (2026). Multi-encoder ConvNeXt network with smooth attentional feature fusion for multispectral semantic segmentation. Neurocomputing, 133533.

[19] Li, F., Jie, J., & Luo, X. (2026, February). Spatial Attention-Guided Multi-Scale Fusion for EfficientNet-Based Fine-Grained Vegetable Classification. In 2026 14th International Conference on Intelligent Control and Information Processing (ICICIP) (pp. 218-225). IEEE.

[20] Lin, H. (2025, October). AGFPNet: A Fine-Grained Classification Model for Food Images Combining Attention Guidance and Feature Pyramids. In 2025 IEEE 7th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) (pp. 47-54). IEEE.

July-September-Issue, Vol. 1, No. 2 (2026) | DOI: 10.66261/nz54mw97

[21] Chen, Z., Wang, J., & Wang, Y. (2025). Enhancing food image recognition by multi-level fusion and the attention mechanism. Foods, 14(3), 461.

[22] Nargesi, M. H., Parian, J. A., & Kheiralipour, K. (2025). Detection of wheat, chickpea, and sea foam in black pepper using hyperspectral imaging technique. Applied Food Research, 5(1), 101031.

[23] Kapetas, D., Kalogeropoulou, E., Christakakis, P., Klaridopoulos, C., & Pechlivani, E. M. (2025). Comparative Evaluation of AI-Based Multi-Spectral Imaging and PCR-Based Assays for Early Detection of Botrytis cinerea Infection on Pepper Plants. Agriculture, 15(2), 164.