

Research paper

MNFF: Multi-nonlinear feature fusion framework for edible insect quality classification

Hye-Young Wang^a, Seoung-Ho Choi^{b,c,*}^a Department of AI Application, Hansung University, 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul, 02876, Republic of Korea^b Faculty of Basic Liberal Art, Hansung University, 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul, 02876, Republic of Korea^c Department of Biomedical Informatics, Korea University College of Medicine, 73, Goryeo-daero, Seongbuk-gu, Seoul, 02841, Republic of Korea

ARTICLE INFO

Keywords:

Edible insect

Explainable AI (XAI)

MambaVision

Multi-nonlinear feature fusion

Quality classification

ABSTRACT

Rapid industrialization of edible insect production and the growing demand for sustainable protein sources have increased the need for reliable quality inspection in large-scale breeding environments. However, early abnormal larvae often exhibit only weak and spatially dispersed morphological differences, making them difficult to detect using conventional vision models. To address this challenge, we propose a Multi-Nonlinear Feature Fusion (MNFF) framework for binary edible insect quality classification and instantiate it as MV-MNFF by integrating the framework with a MambaVision backbone, a convolutional block attention module (CBAM), and a classification head. Unlike conventional attention-based enhancement or simple fusion strategies that primarily reweight or aggregate features, the proposed MNFF framework explicitly models interactions among weak morphological cues by multiplicatively integrating GELU, ELU, and ReLU responses within depthwise separable and multi-scale dilated convolution stages, together with channel attention and skip connections. Experiments on mealworm larvae (*Tenebrio molitor*) and flower chafer larvae (*Protaetia brevitarsis*) showed that the proposed model improved balanced accuracy by 1.96% and 0.40%, respectively, over the baseline MambaVision, while achieving an area under the receiver operating characteristic curve (AUROC) of 1.000 on both datasets. Explainability was further examined using Grad-CAM, LIME, SHAP, and Saliency Map and quantitatively evaluated with faithfulness metrics, including average drop (%), percentage increase in confidence, positive perturbation AUC, and negative perturbation AUC. These results indicate that the proposed MNFF framework and its MambaVision-based implementation provide an effective and explainable solution for early abnormality screening and support more reliable field-level decision-making in industrial-scale edible insect farming.

1. Introduction

As global population growth and increasing food demand intensify the need for sustainable protein sources, edible insects have attracted growing attention owing to their high nutritional value and low environmental burden [1,2]. With the rapid expansion of the edible insect industry, large-scale breeding facilities have become increasingly common, and the importance of systematic health monitoring and quality management has grown accordingly to ensure productivity, food safety, and stable operation [3–6]. However, quality inspection in industrial breeding environments still relies heavily on visual examination or microscopic observation by trained personnel [4,7]. Such approaches are labor- and time-intensive and are further limited by subjective variability among evaluators, making it difficult to ensure consistency and scalability [8]. In large-scale breeding environments where thousands of individuals must be handled simultaneously, it is particularly diffi-

cult to reliably distinguish normal from abnormal individuals, especially when abnormality appears only through subtle morphological changes or weak visual cues. Delayed removal of diseased or abnormal individuals may rapidly accelerate contamination and lead to substantial production losses, thereby increasing the need for reliable automated quality classification systems in industrial settings [9].

To address these challenges, recent studies have introduced computer vision and deep learning-based automatic monitoring systems for tasks such as counting, size measurement, growth prediction, sex determination, and disease detection [7,10,11]. However, most existing insect imaging studies have been designed around relatively salient visual characteristics or clearly observable abnormalities, while the problem of precisely distinguishing normal and abnormal individuals within the same species has remained comparatively underexplored [12,13]. In particular, early signs of quality degradation, such as subtle epidermal changes, slight color variation, localized damage, and mild

* Corresponding author.

E-mail addresses: ladder37@naver.com (H.-Y. Wang), jcn99250@naver.com (S.-H. Choi).<https://doi.org/10.1016/j.rineng.2026.110313>

Received 1 February 2026; Received in revised form 26 March 2026; Accepted 26 March 2026

Available online 30 March 2026

2590-1230/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

morphological abnormalities, are visually weak and spatially dispersed, making them easy to overlook [7,14]. Therefore, the central gap in edible insect quality classification lies not merely in the limited exploration of a new application domain, but in the methodological limitation that feature learning structures capable of reliably representing and integrating weak and heterogeneous visual cues have not yet been sufficiently established.

These limitations are also closely related to the way existing vision architectures form feature representations. Standard convolutional neural networks form hierarchical representations through local receptive field-based feature extraction and repeated downsampling; while effective for summarizing global context, such architectures may allow weak and locally distributed subtle cues to become diluted or overshadowed by stronger responses in deeper layers [15,16]. Attention-based enhancement can be effective for selectively emphasizing important regions or channels and refining feature representations, but such reweighting alone may be insufficient to capture inter-channel correlations and complex interactions among features; therefore, additional nonlinear modeling may be required in problems such as subtle abnormality detection, where combinations of weak visual cues are critical [17,18]. In addition, although recent transformer-based or state-space-based backbones have improved the modeling of long-range dependencies and global context, global information alone may still be insufficient when subtle visual differences are crucial, and the effective integration of complementary features captured by different backbones remains important [19,20]. In other words, for problems characterized by subtle visual differences, simply scaling up the backbone may not be sufficient, and a nonlinear feature fusion structure capable of effectively reflecting interactions among weak local cues is required [21–23].

Edible insect quality assessment thus has the characteristics of a fine-grained visual discrimination problem, in which the information required for classification is formed not by a single salient feature but by a combination of multiple weak visual cues [15,24]. In real industrial environments, such cues may appear as slight discoloration, texture variation, differences in surface gloss, localized damage, and morphological imbalance. Accordingly, to achieve reliable quality assessment in practice, a feature fusion mechanism is required that goes beyond simple linear combination or attention-based emphasis and can jointly reflect multiple nonlinear responses and multi-scale feature interactions. Furthermore, if automated quality classification results are to be used for downstream decisions such as removal, sorting, and monitoring, they must provide not only high predictive accuracy but also visually explainable decision evidence whose reliability can be verified.

Accordingly, the main contributions of this study are as follows:

- We reformulate edible insect quality inspection as a fine-grained binary classification problem in which reliable discrimination depends on weak and spatially dispersed morphological cues. We further identify insufficient modeling of interactions among such cues as a key methodological limitation of existing vision architectures.
- We propose a Multi-Nonlinear Feature Fusion framework for edible insect quality classification and instantiate it as MV-MNFF using a MambaVision backbone. The proposed design organizes post-backbone processing into feature refinement, nonlinear interaction modeling, and decision-oriented representation learning.
- Within this framework, we introduce an MNFF module that performs two-stage multi-nonlinearity-based feature integration. Specifically, GELU, ELU, and ReLU responses are generated in parallel after depth-wise separable and multi-scale dilated convolution stages and are combined through multiplicative fusion. Channel attention and skip connections are further used to refine informative features and stabilize optimization.
- We provide a comprehensive validation of the proposed method through comparisons with standard deep learning backbones and a prior-work baseline, together with component-wise analysis, ablation study, fusion-operation analysis, robustness evaluation under

class-ratio shifts and noise perturbations, and quantitative faithfulness assessment of XAI explanations.

The rest of the paper is organized as follows: Section 2 presents the related works; Section 3 describes the methodology, including the dataset, preprocessing, the proposed method, and the experimental configuration; Section 4 presents the experimental results; Section 5 discusses the findings and limitations of the proposed framework; and Section 6 concludes the paper.

2. Related works

2.1. Deep learning for insect detection and classification

Research applying deep learning to insect recognition has expanded steadily over the past decade. Høye et al.[25] reported that deep learning and computer vision have the potential to substantially advance entomological research by enabling the non-invasive estimation of insect abundance, biomass, and phenotypic diversity through image-based automatic monitoring. Teixeira et al.[26] also systematically reviewed deep learning-based automatic insect detection studies and identified class imbalance, incomplete annotation, the difficulty of small-object recognition, and the discrimination of morphologically similar categories as major challenges. These issues may become even more severe in edible insect production environments, where high-density breeding, occlusion, illumination variation, and high visual similarity among individuals occur simultaneously [17,26].

To mitigate these challenges, prior studies have employed multi-view information, diverse feature combinations, attention mechanisms, and lightweight backbones [17]. For example, An et al.[20] proposed a complementary feature fusion architecture that combines a CNN-based backbone with a transformer-based backbone, while Wu et al.[27] showed through the IP102 benchmark that inter-class similarity and intra-class variation are key challenges in insect recognition. Ren et al.[28] and Liu et al.[29] also sought to improve insect pest recognition performance through a feature reuse structure and a multi-branch feature fusion structure, respectively. However, most of these studies focused on species-level classification or tasks in which visual differences were relatively obvious [17].

Studies directly addressing edible insects remain limited [7]. Baur et al.[30] proposed a non-invasive monitoring system for estimating the size distribution of *Tenebrio molitor* larvae, and Pratondo and Bramantoro[31] showed that transfer learning can be effective for classifying edible insect larvae with similar appearances. Although not directly identical to the edible insect domain, similar attempts have also been reported in sericulture. Mungase et al.[32] compared ResNet-18 and ResNet-50 for early silkworm disease prediction and reported that ResNet-18 achieved superior classification performance. However, that study focused on disease-centered binary classification between diseased and undiseased individuals and therefore did not comprehensively address more diverse abnormal conditions beyond disease-related signs, such as subtle morphological abnormalities, surface damage, and quality degradation. Overall, these studies have demonstrated the feasibility of applying deep learning in edible insect and adjacent domains, but the problem of precisely capturing weak and heterogeneous abnormal cues within the same species remains insufficiently explored [7,17].

2.2. Fine-grained visual classification and the challenge of recognizing subtle visual differences

Fine-grained visual classification addresses the task of distinguishing subcategories in settings where inter-class differences are small and intra-class variation is large. Wei et al.[15] comprehensively reviewed this field and identified discriminative part localization and effective feature representation learning as its core challenges. In particular, the difficulty of fine-grained classification arises from the fact that category

boundaries are determined less by global appearance differences than by combinations of weak local cues such as texture patterns, color changes, and micro-structures [15]. This characteristic is directly relevant to edible insect quality inspection. In practice, early abnormality signs often appear not as a single obvious defect but as weak and spatially distributed signals, such as slight discoloration, irregular surface texture, localized damage, and body-shape imbalance.

Similar phenomena have been reported in fine-grained classification problems in other agricultural domains. Yang et al. [16] showed that conventional convolutional neural network features may fail to sufficiently preserve disease-specific texture and color cues in crop disease classification, and thus argued that structures emphasizing important local regions are needed. Sun et al. [24] pointed out that existing fine-grained classification methods often assume that the most discriminative information will naturally emerge during training, whereas in practice models may become biased toward more salient differences and fail to adequately resolve confusion between highly similar categories. Rao et al. [33] further showed that guiding the model to learn more reliable attention regions can improve its ability to capture subtle inter-class differences.

These findings suggest that edible insect quality inspection cannot be solved simply by using a stronger backbone. The key is not to identify a single dominant feature, but to preserve and enhance multiple weak cues that become meaningful only when they appear together across different representational levels and spatial ranges. Therefore, in subtle abnormality detection, not only backbone capacity but also the way features are integrated and emphasized becomes critical.

2.3. Feature fusion strategies and the need for nonlinear interaction modeling

Feature fusion is a core component of modern deep learning models, and the representational capacity of a model depends strongly on how fusion is performed. Common fusion strategies in insect and agricultural vision studies include relatively simple linear combinations such as concatenation, element-wise addition, and weighted summation [20,34–36]. These methods are computationally efficient and simple to implement, but they may be insufficient when the discriminative signal is formed through the interaction of multiple weak visual cues. Lin et al. [37] pointed out that in nondestructive food quality inspection, quality-related signals often exhibit high complexity and nonlinearity, requiring representation learning that can sufficiently reflect interactions among multiple features. Prunella et al. [21] also emphasized that, in subtle defect detection, strengthening representational capacity is more important than simply increasing model scale.

Accordingly, various attempts have been made in fine-grained recognition to incorporate nonlinear feature interactions. Lin et al. [23] explicitly modeled feature interactions through bilinear CNNs, and Tan et al. [38] proposed hierarchical bilinear pooling to more effectively capture multi-level feature relationships. Zhang et al. [18] further argued that simple linear combinations are insufficient to represent complex feature dependencies and highlighted the need for nonlinear feature fusion structures. These approaches collectively suggest that nonlinear fusion can provide richer representational capacity than linear aggregation.

However, such studies have not directly addressed edible insect quality inspection. Existing insect-domain studies have largely focused on multi-view or complementary feature combinations [20], machine-vision monitoring for edible insect production [7,39], lightweight classifier design [34,40], or attention-based enhancement [17,41], while only limited work has structurally designed nonlinear feature interactions specifically for subtle abnormality detection [11]. Moreover, although attention-based enhancement can be effective for selectively emphasizing important regions or channels, such reweighting alone may be insufficient to fully capture inter-channel correlations and complex feature interactions [17]. Because edible insect quality assessment must simul-

taneously consider weak color variation, subtle texture changes, and localized morphological abnormalities [11,39], simple linear fusion or attention-based emphasis alone may be inadequate for representing the subtle decision boundaries required in this problem [18,23,38]. Therefore, beyond merely preserving feature presence, this problem requires a nonlinear fusion strategy that can strengthen interactions among features.

2.4. Research gaps and motivation for the proposed framework

Taken together, the existing literature reveals three common limitations that must be addressed for edible insect quality inspection. First, studies that precisely distinguish normal and abnormal individuals within the same species remain limited in the edible insect domain, while most prior work has focused on species identification, counting, population monitoring, or the detection of visually explicit abnormal signs [25,26,30,35,42]. Although the IP102 benchmark introduced by Wu et al. [27] provided an important methodological foundation for insect recognition, a representation learning structure capable of effectively capturing subtle and heterogeneous abnormal cues, including early abnormalities, has not yet been sufficiently established in the edible insect domain.

Second, early abnormality signs in edible insects tend to appear not as a single salient defect but as combinations of multiple weak visual cues, including slight discoloration, texture variation, changes in surface gloss, localized damage, and morphological imbalance. This implies that the discriminative signal is likely formed not by an individual feature alone but by interactions among multiple cues [15,16,24]. Therefore, linear fusion that treats each feature contribution as independent or purely additive, as well as attention-based emphasis that mainly recalibrates feature importance, may be insufficient to adequately highlight subtle abnormality-related patterns [17,18,37].

Third, although existing nonlinear fusion methods have shown utility in fine-grained recognition, they have not directly addressed problems such as edible insect quality inspection, where visual cues across different spatial scales and different nonlinear responses must be considered simultaneously [18,23,38,43,44]. In other words, prior studies have partially demonstrated the importance of feature integration in the insect domain, but they have not sufficiently proposed structures that jointly incorporate multiple nonlinear responses, multi-scale features, and channel-wise refinement for subtle abnormality detection [7,20,34,35].

Overall, edible insect quality assessment is a fine-grained visual discrimination problem whose core challenge lies in representation learning that can effectively amplify and preserve weak visual cues that become meaningful only when they appear together [15,24,33]. Therefore, linear fusion or attention-based emphasis alone is not sufficient, and a strengthened feature fusion structure capable of jointly reflecting multiple nonlinear responses and multi-scale feature interactions is required [21–23].

3. Methodology

3.1. Dataset and data preprocessing

Dataset Configuration: The dataset consists of images of two representative edible insect species, mealworm larvae (*Tenebrio molitor*) and flower chafer larvae (*Protaetia brevitarsis*). Each image contains multiple individuals, and for each species the dataset includes 1000 abnormal images and 1000 normal images, for a total of 2000 images. The task is defined as a binary classification problem, where abnormal is labeled as 1 and normal as 0. The data can be obtained through the link provided in Section 6.

Data Preprocessing: To preserve the inherent visual characteristics of real farm environments and evaluate robustness under practical conditions, only minimal preprocessing was applied. All images were loaded

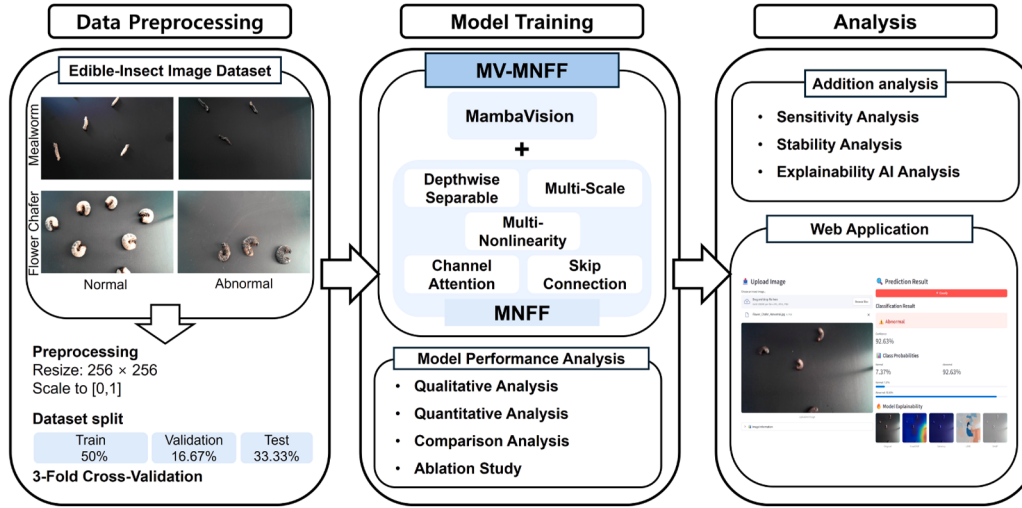


Fig. 1. Overview of framework.

using OpenCV, and images read in BGR format were converted to RGB during loading. Then, all images were resized to 256×256 to ensure input size consistency, and pixel values were normalized to the range of $[0, 1]$ by dividing by 255. Finally, the channel order was converted from HWC to CHW to match the PyTorch input format, and the resulting tensors were used for model training and evaluation. No additional preprocessing, such as data augmentation, color correction, denoising, or histogram equalization, was applied. This setting was designed to evaluate whether the proposed MNFF framework can directly extract discriminative features from raw images captured in real breeding environments. For fair comparison, the same preprocessing procedure was identically applied to all baseline and comparative models Fig. 1.

3.2. Proposed method

Overall Framework: The proposed framework provides an end-to-end pipeline for edible insect quality classification, encompassing data preprocessing, model training, and deployment in Fig. 1. Input images of mealworm and flower chafer larvae are resized and normalized, then binary-labeled as Normal or Abnormal. A MambaVision-based classifier is enhanced with the proposed multi-nonlinear feature fusion (MNFF) framework, which integrates CBAM, multi-scale and depthwise separable convolutions with multiple nonlinear activations, channel attention, and skip connections to capture fine-grained morphological variations. Model robustness and reproducibility were assessed using 3-fold cross-validation and bootstrap-based confidence interval estimation. For practical applicability, explainable AI analysis and a web-based application are incorporated, enabling real-time classification with visual explanations to support transparent and reliable decision-making in large-scale insect farming environments.

Architecture and Functional Role of the MNFF Framework with MambaVision: Fig. 2 presents the overall architecture of the proposed MV-MNFF model, which integrates the multi-nonlinear feature fusion (MNFF) framework with a MambaVision backbone. The MNFF framework is designed as a feature enhancement module to amplify subtle morphological cues that are insufficiently captured by conventional backbone representations. Fig. 2a illustrates the internal structure of the MNFF module.

Given an input feature map, depthwise separable convolution is first applied to efficiently extract spatial features while maintaining computational efficiency. Subsequently, a multi-nonlinearity branch processes the feature map in parallel using three activation functions. GELU provides smooth and continuous responses to small input variations, enabling sensitivity to microscopic morphological changes. ELU stabilizes

feature distributions and facilitates deep feature learning by maintaining non-zero gradients in negative regions, while ReLU enforces sparsity by suppressing negative activations and emphasizing salient responses. The resulting nonlinear feature maps are fused via element-wise multiplication to jointly represent diverse nonlinear patterns that are difficult to capture with a single activation function. This multiplication-based fusion is intended to emphasize discriminative regions that are consistently supported across different nonlinear responses while suppressing inconsistent activations that are specific to individual branches.

To further enhance spatial representation, multi-scale dilated convolutions with dilation rates $d \in 1, 2, 4$ are applied independently to each nonlinear feature branch, yielding effective receptive fields of 3×3 , 5×5 , and 9×9 , respectively. This design is intended to help the network capture insects with varying sizes and morphological structures. The multi-scale features are subsequently fused through element-wise multiplication, reinforcing hierarchical representations by jointly modeling nonlinear characteristics across multiple spatial scales.

Channel-wise feature refinement is then performed using Global Average Pooling followed by fully connected layers to model inter-channel dependencies. ReLU activation is applied between the fully connected layers to introduce nonlinearity into the channel weighting process, and also at the output stage to suppress negative channel responses and highlight informative feature channels. Finally, the enhanced features are combined with the original input through a skip connection, preserving both fine-grained local details and global contextual information essential for accurate insect quality classification.

Fig. 2b illustrates the classification head, which consists of three fully connected blocks, each composed of a linear layer, BatchNorm1d, activation function, and dropout. LeakyReLU is employed in the first block to maintain gradient flow for weak features, while ReLU is used in the subsequent blocks. A dropout rate of 0.5 is applied in each block to mitigate overfitting. This progressive abstraction transforms the MNFF-enhanced features into a final binary classification decision.

As shown in Fig. 2c, the MNFF framework is embedded after the final feature extraction stage of the MambaVision backbone. Hierarchical features extracted by MambaVision are first refined using a CBAM block to emphasize salient spatial and channel-wise information. The MNFF module is then applied to capture additional microscopic morphological variations and texture patterns that may be overlooked by the backbone. The resulting features are converted into fixed-length vectors via Global Average Pooling and flattening, and subsequently fed into the classification head to predict Normal or Abnormal insect conditions.

Overall, the MNFF framework serves as a modular feature enhancement mechanism that combines multi-nonlinearity, multi-scale spatial

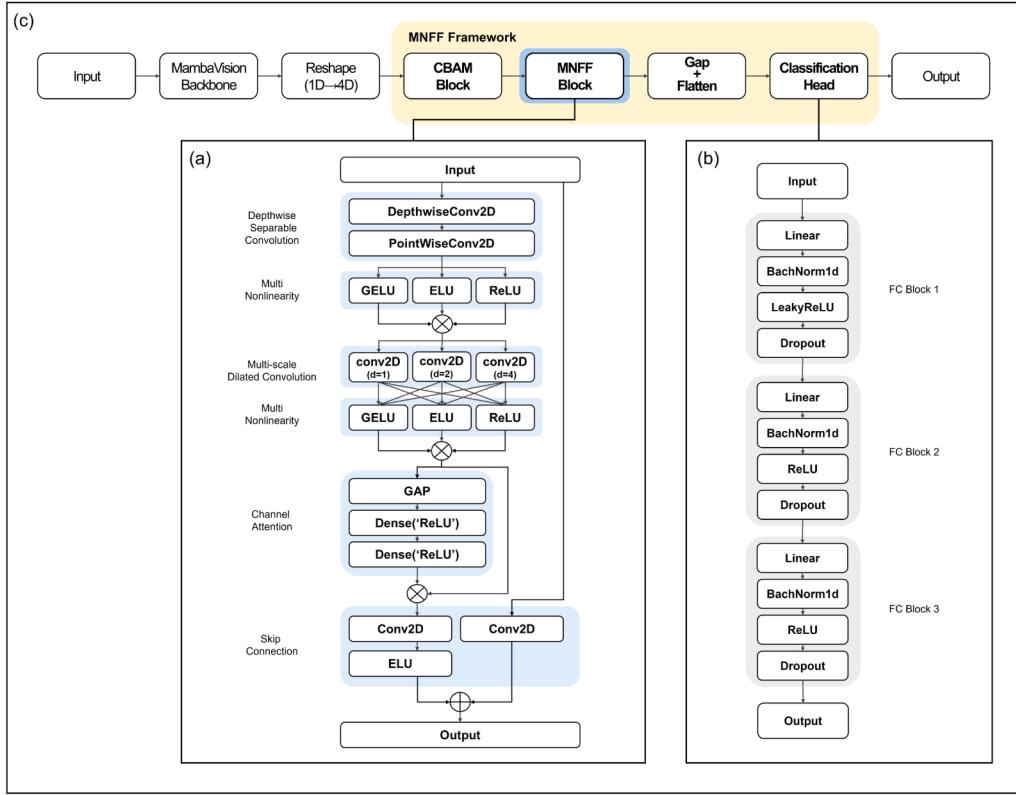


Fig. 2. The overall structure of the proposed MV-MNFF model. (a) The architecture of the MNFF Module, (b) The architecture of the Classification Head, (c) The complete pipeline of MambaVision integrated with MNFF Framework.

modeling, channel attention, and residual learning. When integrated with the MambaVision backbone, it is intended to improve the representation of subtle visual patterns relevant to industrial-scale edible insect quality management.

Algorithm Overview of the MNFF Module: Algorithm 1 shows the entire operational procedure of the MNFF module proposed in this study in pseudocode format. This module consists of four main steps: first, depthwise separable convolution (DSC) with multi-nonlinearity (MN); second, multi-scale dilated convolution (MS) with multi-nonlinearity (MN); third, channel attention (CA); and finally, skip connection (SC). This step-by-step design is intended to extract features at different scales, focus on critical channels, and promote stable gradient flow for preserving microscopic morphological features in insect images. Specifically, by simultaneously utilizing the GELU, ELU, and ReLU activation functions at each step, the module is intended to capture diverse nonlinear properties and combine them through element-wise multiplication.

Phase 1 is a depthwise separable convolution with multi-nonlinearity. To extract spatial features while maintaining computational efficiency, the input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C_{in}}$ first passes through a depthwise separable convolution with three parallel activation functions. It is defined as follows:

$$\mathbf{X}_i = \sigma_i(\text{Conv}_{1 \times 1}(\text{DepthwiseConv}_{3 \times 3}(\mathbf{X}))), \quad i \in \{1, 2, 3\} \quad (1)$$

where $\sigma_1 = \text{GELU}$, $\sigma_2 = \text{ELU}$, and $\sigma_3 = \text{ReLU}$, the three feature maps are then fused by element-by-element multiplication as follows:

$$\mathbf{X}_{fused} = \mathbf{X}_1 \odot \mathbf{X}_2 \odot \mathbf{X}_3 \quad (2)$$

In Eq. (2), $\odot$ represents element-wise multiplication. In Eq. (1), depthwise separable convolution can efficiently extract spatial features while reducing the number of parameters and computational cost compared with standard convolution. Meanwhile, unlike additive fusion, the multiplicative fusion in Eq. (2) mitigates the tendency for the fused response to be dominated by a large activation from a single branch and exhibits

Algorithm 1 Multi-nonlinearity feature fusion module.

Require: Input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C_{in}}$

Require: Output channels C_{out}

Ensure: Enhanced feature map $\mathbf{Y} \in \mathbb{R}^{H \times W \times C_{out}}$

- 1: **Phase 1: Depthwise Separable Convolution with Multi-Nonlinearity**
 - 2: $\mathbf{X}_{dw} \leftarrow \text{DepthwiseConv2D}(\mathbf{X}, k = 3, \text{padding} = \text{'same'})$
 - 3: $\mathbf{X}_{pw} \leftarrow \text{PointwiseConv2D}(\mathbf{X}_{dw}, C_{out}, k = 1)$
 - 4: $\mathbf{X}_1 \leftarrow \text{GELU}(\mathbf{X}_{pw})$
 - 5: $\mathbf{X}_2 \leftarrow \text{ELU}(\mathbf{X}_{pw})$
 - 6: $\mathbf{X}_3 \leftarrow \text{ReLU}(\mathbf{X}_{pw})$
 - 7: $\mathbf{X}_{fused} \leftarrow \mathbf{X}_1 \odot \mathbf{X}_2 \odot \mathbf{X}_3$
 - 8: **Phase 2: Multi-Scale Dilated Convolutions with Multi-Nonlinearity**
 - 9: **for** $\sigma \in \{\text{GELU}, \text{ELU}, \text{ReLU}\}$ **do**
 - 10: $\mathbf{D}_1^\sigma \leftarrow \sigma(\text{Conv2D}(\mathbf{X}_{fused}, C_{out}, k = 3, d = 1))$
 - 11: $\mathbf{D}_2^\sigma \leftarrow \sigma(\text{Conv2D}(\mathbf{X}_{fused}, C_{out}, k = 3, d = 2))$
 - 12: $\mathbf{D}_4^\sigma \leftarrow \sigma(\text{Conv2D}(\mathbf{X}_{fused}, C_{out}, k = 3, d = 4))$
 - 13: $\mathbf{F}^\sigma \leftarrow \text{Concatenate}([\mathbf{D}_1^\sigma, \mathbf{D}_2^\sigma, \mathbf{D}_4^\sigma])$
 - 14: **end for**
 - 15: $\mathbf{F}_{multi} \leftarrow \mathbf{F}^{\text{GELU}} \odot \mathbf{F}^{\text{ELU}} \odot \mathbf{F}^{\text{ReLU}} \in \mathbb{R}^{H \times W \times 3C_{out}}$
 - 16: **Phase 3: Channel Attention**
 - 17: $\mathbf{g} \leftarrow \text{GlobalAveragePooling2D}(\mathbf{F}_{multi}) \in \mathbb{R}^{3C_{out}}$
 - 18: $\mathbf{s}_1 \leftarrow \text{ReLU}(\text{Dense}(\mathbf{g}, C_{out}))$
 - 19: $\mathbf{s}_2 \leftarrow \text{ReLU}(\text{Dense}(\mathbf{s}_1, 3C_{out}))$
 - 20: $\mathbf{A} \leftarrow \text{Reshape}(\mathbf{s}_2, (1, 3C_{out}, 1, 1))$ {Channel attention weights}
 - 21: $\mathbf{F}_{att} \leftarrow \mathbf{F}_{multi} \times \mathbf{A}$ {Channel-wise scaling}
 - 22: **Phase 4: Skip Connection**
 - 23: $\mathbf{F}_{fused} \leftarrow \text{ELU}(\text{Conv2D}(\mathbf{F}_{att}, C_{out}, k = 3, \text{padding} = \text{'same'}))$
 - 24: $\mathbf{X}_{skip} \leftarrow \text{Conv2D}(\mathbf{X}, C_{out}, k = 1)$
 - 25: $\mathbf{Y} \leftarrow \mathbf{F}_{fused} + \mathbf{X}_{skip}$
 - 26: **return** $\mathbf{Y}$
-

a soft AND-type gating property that relatively emphasizes locations where multiple nonlinearity branches simultaneously produce high responses. This property is advantageous for suppressing noise-like activations and inducing more stable feature representations when the target patterns are weak, as in micro-morphological cues.

In addition, gradient propagation through multiplicative fusion can be derived according to the product rule. Let the shared pre-activation feature map be denoted by u , and define $x_1 = \text{GELU}(u)$, $x_2 = \text{ELU}(u)$, and $x_3 = \text{ReLU}(u)$. The fused representation is then given by $x_{\text{fused}} = x_1 \odot x_2 \odot x_3$:

$$\frac{\partial x_{\text{fused}}}{\partial u} = x_2 \odot x_3 \odot \text{GELU}'(u) + x_1 \odot x_3 \odot \text{ELU}'(u) + x_1 \odot x_2 \odot \text{ReLU}'(u). \quad (3)$$

Eq. (3) shows that the gradient of each branch is scaled by the output values of the other branches, suggesting that multiplicative fusion can function as a soft-gating mechanism that reflects the agreement among branch responses. At the same time, when the output of one or more branches becomes close to zero at certain locations, the product structure may cause gradient attenuation, which can potentially lead to optimization instability. To alleviate this issue, the proposed module provides an identity path through the skip connection in Eq. (9), helping preserve the overall gradient flow even when gradients along the multiplicative path become weak and thereby improving training stability.

Phase 2 is the multi-scale dilated convolution with multi-nonlinearity. Three parallel dilated convolutions with expansion coefficients of 1, 2, and 4 were applied for each activation function to capture features at different scales:

$$\mathbf{F}^\sigma = \text{Concat}[\text{Conv}_{d=1}^\sigma(\mathbf{X}_{\text{fused}}), \text{Conv}_{d=2}^\sigma(\mathbf{X}_{\text{fused}}), \text{Conv}_{d=4}^\sigma(\mathbf{X}_{\text{fused}})] \quad (4)$$

In Eq. (4), $\sigma \in \{\text{GELU}, \text{ELU}, \text{ReLU}\}$, we then fuse the three multi-scale feature maps into element-by-element

$$\mathbf{F}_{\text{multi}} = \mathbf{F}^{\text{GELU}} \odot \mathbf{F}^{\text{ELU}} \odot \mathbf{F}^{\text{ReLU}} \quad (5)$$

The connected outputs of Eqs. (4) and (5) are intended to incorporate information from multiple receptive fields to form richer spatial context.

Phase 3 is Channel Attention. To highlight the important channels, a channel attention mechanism based on Squeeze-and-Excitation (SE) is applied. First, a GAP is performed for $\mathbf{F}_{\text{multi}}$. Then, the global information of each channel is compressed as follows:

$$\mathbf{g} = \text{GAP}(\mathbf{F}_{\text{multi}}), \quad \mathbf{g} \in \mathbb{R}^{3C_{\text{out}}} \quad (6)$$

We then model the inter-channel dependence through two fully connected layers:

$$\mathbf{s}_1 = \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{g}), \quad \mathbf{s}_2 = \text{ReLU}(\mathbf{W}_2 \cdot \mathbf{s}_1) \quad (7)$$

Where $\mathbf{W}_1 \in \mathbb{R}^{C_{\text{out}} \times 3C_{\text{out}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{3C_{\text{out}} \times C_{\text{out}}}$ which is restored after dimensionality reduction. We highlight the important channel by multiplying the finally learned attention weight $\mathbf{A}$ by $\mathbf{F}_{\text{multi}}$:

$$\mathbf{F}_{\text{att}} = \mathbf{F}_{\text{multi}} \times \text{Reshape}(\mathbf{s}_2, (1, 3C_{\text{out}}, 1, 1)) \quad (8)$$

Phase 4 is the Skip Connection. Finally, a skip connection is applied to improve learning stability and gradient flow. It fuses the attention-applied feature map and adds the original input to generate the final output as follows:

$$\mathbf{Y} = \text{ELU}(\text{Conv}_{3 \times 3}(\mathbf{F}_{\text{att}})) + \text{Conv}_{1 \times 1}(\mathbf{X}) \quad (9)$$

The skip connection in Eq. (9) combines the low-level spatial information of the original input with the high-level semantic representation that has undergone multi-layered transformation. This design is intended to support feature representation for classification tasks that require precise structural preservation, such as the microscopic morphological features of insects. In addition, the skip connection provides an identity path that helps alleviate gradient attenuation that may arise

from multiplicative fusion, thereby supporting optimization stability and overall gradient flow.

Loss Function The proposed model is optimized to minimize Cross-Entropy loss in the learning process. Cross-Entropy loss quantifies the difference between predictive probability and actual labels and is defined as follows:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{p}_{i,c}), \quad (10)$$

In Eq. (10) N is the number of samples, C is the number of classes, $y_{i,c}$ is the one-hot real label, and $\hat{p}_{i,c}$ is the predictive probability.

3.3. Experimental configuration and training protocol

Data Splitting and Cross-Validation Strategy: To objectively evaluate the generalization performance of the model, we first applied random shuffling to prevent the sample order from being biased toward a specific class, and then adopted a stratified 3-fold cross-validation strategy. The entire dataset was first divided into three folds of equal size, and stratified sampling was applied so that the class ratio remained balanced within each fold. In each iteration, one fold, corresponding to 33.33% of the data, was designated as an independent test set that was completely excluded from the training process, while the remaining two folds, corresponding to 66.67%, were combined to form the development set. This development set was then split by random sampling into 75% for the training set and 25% for the validation set while preserving the class ratio. As a result, each experiment was conducted with an overall ratio of 50% training, 16.67% validation, and 33.33% test data. This ratio was not manually predetermined but resulted from the stratified 3-fold cross-validation design. The relatively large test portion was intentionally retained so that each iteration could be evaluated on a fully held-out fold with a sufficient amount of unseen data. The effective training portion became 50% because the remaining development set was further divided to preserve an independent validation set for model selection and early stopping, rather than maximizing the training size at the expense of evaluation independence. The test fold was rotated across the three iterations, and the final performance was obtained by averaging the test results from all iterations.

Classification Models: The hyperparameter settings of the classification models used in this study are summarized in Table 1. We employed MambaVision [19], NextViT [46], ResNet [47], DenseNet [48], and EfficientNetV2 [45] as classification models. For MambaVision and EfficientNetV2, additional configurations were constructed by integrating the MNFF framework into each original backbone, namely MV-MNFF (Our Proposal 1) and EN-MNFF (Our Proposal 2), and their performance was evaluated. In addition, ResNet-18 [47], which showed the best performance among the comparative models in the prior study of Mungase et al. [32], was adopted as the prior-work baseline.

Training Strategy: All models were trained using the Adam optimizer and Cross-Entropy loss with a batch size of 32 and a learning rate of 1×10^{-3} . The learning rate was kept constant throughout training to examine stable convergence behavior, and training was conducted for up to 50 epochs. In addition, early stopping with a patience of 10 was applied such that training was terminated when validation accuracy did not improve for 10 consecutive epochs, and the weights corresponding to the best validation performance were restored. Meanwhile, the prior-work baseline ResNet-18 was trained by following the model configuration of the original study, using ImageNet pretrained weights, freezing the convolutional layers, and fine-tuning only the final fully connected layer.

Hardware and Implementation Details: All experiments were conducted using the PyTorch framework on an NVIDIA RTX A6000 GPU. Except for the prior-work baseline ResNet-18, all models were trained from scratch without ImageNet pretrained weights in order to evaluate their intrinsic feature learning capability, and the default PyTorch initialization scheme was applied to all network components including the

Table 1
Hyperparameter settings for the experimental models.

MambaVision-Tiny [19]		EfficientNetV2-S [45]	
Dimension	80	Backbone dropout	0.2
Input dimension	32	Stochastic depth	0.2
Depths	[1,3,8,4]	Hidden dimension	512
Number of heads	[2,4,8,16]	Classifier dropout	0.1
Window size	[8,8,14,7]		
MLP ratio	4		
Drop path rate	0.2		
Hidden dimension	512		
Classifier dropout	0.1		
NextViT-Small [46]		ResNet-50 [47]	
Stem channels	[64,32,64]	Feature dimension	2048
Depths	[3,4,10,3]	FC layers	[512,512,256]
Path dropout	0.1	FC stack dropout	0.5
Hidden dimension	512		
Classifier dropout	0.1		
DenseNet-121 [48]			
Feature dimension	1024		
FC layers	[512,512,256]		
FC stack dropout	0.5		
MV-MNFF (Our Prop. 1) (Backbone: MambaVision-Tiny)		EN-MNFF (Our Prop. 2) (Backbone: EfficientNetV2-S)	
Dimension	80	Backbone dropout	0.2
Input dimension	32	Stochastic depth	0.2
Depths	[1,3,8,4]	Hidden dimension	512
Number of heads	[2,4,8,16]	Classifier dropout	0.1
Window size	[8,8,14,7]		
MLP ratio	4		
Drop path rate	0.2		
Hidden dimension	512		
FC stack dropout	0.5		

proposed MNFF module. In addition, the random seed was fixed at 1004 to ensure reproducibility.

Classification Performance Metrics: Balanced accuracy (BA) was adopted as the primary evaluation metric to ensure fair performance assessment under class-imbalanced conditions, as it equally reflects sensitivity and specificity. To quantify statistical reliability, 95% confidence intervals for all metrics were estimated using 1000 bootstrap resamplings. In addition to BA, model performance was comprehensively evaluated using Loss, Precision, Recall, F1 score, AUROC, and AUPR. These metrics were derived from the confusion matrix components, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). AUROC was used to assess threshold-independent discriminative ability based on the trade-off between true positive and false positive rates, while AUPR provided a more informative measure of positive-class prediction performance, particularly under severe class imbalance.

Faithfulness Metrics for Visual Explanations: To quantitatively evaluate the faithfulness of visual explanations generated by XAI methods, we adopted two types of perturbation-based metrics. First, average drop% and %increase in confidence proposed by Chattopadhyay et al. [49] were used to measure confidence changes under single-step masking. However, because such single-step masking metrics can vary depending on mask generation, threshold selection, and mask area, we additionally reported the positive and negative perturbation AUC proposed by Chefer et al. [50]. Perturbation AUC evaluates the confidence change curve by progressively inserting or removing input regions according to explanation importance, thereby providing a complementary perspective that reduces sensitivity to a single masking configuration and more robustly assesses explanation faithfulness.

By jointly considering these complementary metrics, the evaluation captures classification accuracy, robustness to class imbalance, reliability across decision thresholds, and the faithfulness of visual explanations.

Formal definitions of all evaluation metrics, including both classification performance metrics and faithfulness metrics, are provided in Section 1 of the Supplementary Material.

4. Experimental results

4.1. Overall classification performance

Classification Performance Evaluation: Table 2 summarizes the averaged results of 3-fold cross-validation for multiple deep learning models on the mealworm and flower chafer datasets, comparing the proposed MNFF-integrated models (MV-MNFF and EN-MNFF) with their corresponding baseline backbones. Detailed per-fold performance results are provided in the Supplementary Table 2. For mealworms, all baseline models achieved high performance with BA values above 0.970, where EfficientNetV2 showed the highest BA of 0.987. However, ResNet-50 and DenseNet-121 exhibited relatively higher loss values, indicating reduced predictive stability. After integrating the MNFF framework, MambaVision demonstrated a clear performance gain, with BA increasing from 0.970 to 0.989 and average loss decreasing by more than 50%, whereas EfficientNetV2 showed a marginal BA decrease while maintaining competitive performance.

For flower chafers, baseline models again showed strong performance, with ResNet-50 achieving the highest BA of 0.996. Similar to the mealworm results, relatively higher loss values were observed for ResNet-50 and DenseNet-121. With MNFF integration, MambaVision improved BA from 0.992 to 0.996 and reduced loss by approximately 34%, while EfficientNetV2 maintained its baseline BA. Across both datasets, the MV-MNFF model consistently achieved an AUROC of 1.000, indicating excellent discriminative capability. Overall, these results demonstrate that the MNFF framework effectively enhances classification accuracy and predictive stability when combined with MambaVision, supporting the suitability of MV-MNFF for robust edible insect quality assessment in industrial agricultural settings.

4.2. MNFF module analysis

Component-Wise Analysis: Table 3 reports a component-wise evaluation of the MNFF module, assessing the individual contributions of DSC with MN, MS with MN, channel attention (CA), and skip connection (SC) within MV-MNFF and EN-MNFF using 3-fold cross-validation. Only the averaged performance across folds is reported in Table 3, while detailed per-fold results are provided in Supplementary Table 3. For the mealworm dataset, MV-MNFF showed consistent trends across folds, where CA achieved the best overall performance, yielding the highest BA (average 0.996) and the lowest loss, while all other components also maintained strong performance with BA values above 0.918. In contrast, EN-MNFF benefited most from DSC with MN, which achieved the highest average BA of 0.995, although SC showed fold-dependent superiority in one split.

For the flower chafer dataset, MV-MNFF primarily benefited from DSC with MN, achieving the best average BA, with minor fold-level variations where SC marginally outperformed in one fold. EN-MNFF exhibited more diverse optimal configurations, where MS with MN and SC achieved the highest average BA of 0.998, highlighting the importance of multi-scale feature extraction and residual information preservation. Across all configurations and species, AUROC values consistently exceeded 0.976, indicating reliable discriminative performance. Overall, the results demonstrate that while the optimal MNFF component depends on both the backbone architecture and insect species, multi-scale and multi-nonlinearity-based components consistently contribute strong and robust performance, validating the effectiveness of the MNFF design for edible insect quality classification.

Ablation Study: Table 4 summarizes the ablation analysis of CBAM and MNFF modules. For MV-MNFF, using only one of CBAM or MNFF led to a modest performance drop, whereas the full configuration

Table 2

Performance metrics of classification models across all experimental settings (3-fold average performance).

Model	Balanced Accuracy↑ (95% CI)	Loss↓ (95% CI)	Recall↑ (95% CI)	Precision↑ (95% CI)	F1 Score↑ (95% CI)	AUROC↑ (95% CI)
Mealworm						
MambaVision [19]	0.970(0.957-0.982)	0.066(0.062-0.069)	0.970(0.957-0.982)	0.971(0.959-0.982)	0.970(0.957-0.982)	0.998(0.996-0.999)
EfficientNetV2 [45]	0.987(0.978-0.995)	0.048(0.042-0.053)	0.987(0.978-0.995)	0.987(0.978-0.995)	0.987(0.977-0.994)	0.998(0.996-1.000)
NextViT [46]	0.977(0.965-0.987)	0.060(0.056-0.063)	0.977(0.966-0.987)	0.977(0.965-0.987)	0.977(0.965-0.987)	0.999(0.997-1.000)
ResNet [47]	0.982(0.971-0.991)	0.331(0.329-0.333)	0.982(0.971-0.991)	0.982(0.971-0.991)	0.982(0.971-0.991)	0.997(0.994-1.000)
DenseNet [48]	0.986(0.977-0.995)	0.325(0.324-0.327)	0.986(0.977-0.995)	0.986(0.977-0.995)	0.986(0.977-0.995)	0.999(0.998-1.000)
MV-MNFF (Our Proposal 1)	0.989(0.982-0.997)	0.032(0.030-0.035)	0.989(0.981-0.996)	0.990(0.982-0.997)	0.989(0.981-0.996)	1.000(0.999-1.000)
EN-MNFF (Our Proposal 2)	0.981(0.969-0.991)	0.058(0.053-0.063)	0.981(0.970-0.990)	0.981(0.970-0.991)	0.981(0.969-0.990)	0.999(0.998-1.000)
Flower Chafer						
MambaVision [19]	0.992(0.985-0.997)	0.032(0.027-0.037)	0.992(0.986-0.997)	0.992(0.986-0.997)	0.992(0.985-0.997)	1.000(0.999-1.000)
EfficientNetV2 [45]	0.994(0.988-0.999)	0.018(0.015-0.020)	0.995(0.988-0.999)	0.995(0.989-0.999)	0.994(0.988-0.999)	1.000(1.000-1.000)
NextViT [46]	0.989(0.981-0.997)	0.027(0.024-0.030)	0.990(0.981-0.997)	0.990(0.982-0.997)	0.990(0.982-0.997)	1.000(0.999-1.000)
ResNet [47]	0.996(0.990-1.000)	0.318(0.317-0.319)	0.995(0.990-1.000)	0.996(0.990-1.000)	0.996(0.990-1.000)	1.000(1.000-1.000)
DenseNet [48]	0.993(0.986-0.998)	0.321(0.320-0.322)	0.993(0.986-0.998)	0.993(0.986-0.997)	0.992(0.986-0.998)	1.000(0.999-1.000)
MV-MNFF (Our Proposal 1)	0.996(0.992-0.999)	0.021(0.019-0.023)	0.996(0.992-0.999)	0.996(0.992-0.999)	0.996(0.992-0.999)	1.000(1.000-1.000)
EN-MNFF (Our Proposal 2)	0.994(0.989-0.999)	0.023(0.021-0.025)	0.994(0.989-0.999)	0.995(0.989-0.999)	0.994(0.989-0.998)	1.000(0.999-1.000)

Table 3

Component-wise comparison analysis of MNFF module components (3-fold average performance).

Model	Only Component	BA↑ (95% CI)	Loss↓ (95% CI)	Recall↑ (95% CI)	Precision↑ (95% CI)	F1 Score↑ (95% CI)	AUROC↑ (95% CI)
Mealworm							
MV-MNFF (Our Proposal 1)	DSC with MN	0.989(0.982-0.996)	0.045(0.042-0.047)	0.989(0.981-0.997)	0.990(0.981-0.996)	0.990(0.982-0.996)	0.999(0.997-1.000)
	MS with MN	0.960(0.945-0.973)	0.110(0.105-0.115)	0.960(0.945-0.973)	0.961(0.945-0.973)	0.960(0.945-0.974)	0.992(0.988-0.995)
	CA	0.996(0.991-0.999)	0.021(0.019-0.022)	0.995(0.990-0.999)	0.996(0.990-0.999)	0.995(0.990-0.999)	1.000(0.999-1.000)
	SC	0.988(0.980-0.995)	0.036(0.033-0.040)	0.988(0.981-0.995)	0.989(0.980-0.995)	0.989(0.981-0.995)	0.999(0.999-1.000)
EN-MNFF (Our Proposal 2)	DSC with MN	0.995(0.989-0.999)	0.025(0.023-0.027)	0.995(0.989-0.999)	0.995(0.989-0.999)	0.995(0.989-0.999)	1.000(0.999-1.000)
	MS with MN	0.989(0.980-0.995)	0.031(0.028-0.034)	0.988(0.980-0.995)	0.989(0.980-0.995)	0.988(0.980-0.995)	0.999(0.999-1.000)
	CA	0.986(0.977-0.994)	0.050(0.046-0.055)	0.986(0.977-0.994)	0.986(0.977-0.994)	0.986(0.977-0.994)	0.999(0.997-0.999)
	SC	0.990(0.982-0.995)	0.040(0.038-0.043)	0.990(0.982-0.995)	0.989(0.982-0.995)	0.989(0.982-0.995)	1.000(1.000-1.000)
Flower Chafer							
MV-MNFF (Our Proposal 1)	DSC with MN	0.998(0.994-1.000)	0.009(0.008-0.010)	0.998(0.994-1.000)	0.998(0.995-1.000)	0.998(0.995-1.000)	1.000(1.000-1.000)
	MS with MN	0.994(0.987-0.998)	0.021(0.019-0.023)	0.993(0.988-0.998)	0.993(0.987-0.998)	0.993(0.988-0.998)	0.999(0.998-1.000)
	CA	0.991(0.984-0.998)	0.040(0.036-0.043)	0.991(0.984-0.998)	0.992(0.985-0.998)	0.991(0.984-0.997)	1.000(1.000-1.000)
	SC	0.992(0.985-0.997)	0.030(0.027-0.033)	0.991(0.986-0.997)	0.992(0.986-0.997)	0.991(0.986-0.997)	1.000(1.000-1.000)
EN-MNFF (Our Proposal 2)	DSC with MN	0.995(0.990-0.998)	0.029(0.027-0.031)	0.994(0.989-0.998)	0.995(0.989-0.999)	0.995(0.990-0.998)	1.000(1.000-1.000)
	MS with MN	0.998(0.994-1.000)	0.010(0.008-0.012)	0.998(0.994-1.000)	0.998(0.994-1.000)	0.997(0.994-1.000)	1.000(1.000-1.000)
	CA	0.997(0.993-1.000)	0.013(0.013-0.015)	0.997(0.993-1.000)	0.997(0.993-1.000)	0.997(0.993-1.000)	1.000(1.000-1.000)
	SC	0.998(0.995-0.999)	0.010(0.009-0.012)	0.998(0.995-0.999)	0.997(0.995-0.999)	0.998(0.995-1.000)	1.000(1.000-1.000)

Table 4

Ablation study on the effectiveness of CBAM and MNFF modules (3-Fold average performance).

Model	Variant	BA↑ (95% CI)	Loss↓ (95% CI)	Recall↑ (95% CI)	Precision↑ (95% CI)	F1 Score↑ (95% CI)	AUROC↑ (95% CI)
Mealworm							
MV-MNFF (Our Proposal 1)	Full	0.989(0.981-0.997)	0.032(0.030-0.035)	0.989(0.981-0.996)	0.990(0.981-0.997)	0.989(0.981-0.996)	1.000(0.999-1.000)
	w/o CBAM	0.985(0.975-0.993)	0.045(0.041-0.049)	0.985(0.975-0.994)	0.985(0.975-0.994)	0.985(0.975-0.994)	0.999(0.996-1.000)
	w/o MNFF	0.981(0.970-0.990)	0.048(0.045-0.051)	0.981(0.971-0.990)	0.981(0.971-0.990)	0.981(0.971-0.990)	0.999(0.998-1.000)
	w/o CBAM & MNFF	0.988(0.979-0.996)	0.036(0.033-0.039)	0.988(0.978-0.995)	0.988(0.979-0.996)	0.987(0.979-0.995)	0.999(0.998-1.000)
EN-MNFF (Our Proposal 2)	Full	0.981(0.970-0.990)	0.058(0.053-0.063)	0.981(0.969-0.991)	0.981(0.970-0.990)	0.980(0.970-0.990)	0.999(0.998-1.000)
	w/o CBAM	0.993(0.986-0.998)	0.025(0.023-0.028)	0.993(0.986-0.998)	0.993(0.986-0.998)	0.993(0.986-0.998)	1.000(0.999-1.000)
	w/o MNFF	0.989(0.980-0.996)	0.030(0.027-0.034)	0.989(0.980-0.996)	0.989(0.981-0.996)	0.989(0.980-0.996)	1.000(0.999-1.000)
	w/o CBAM & MNFF	0.984(0.974-0.993)	0.058(0.053-0.064)	0.984(0.974-0.992)	0.984(0.975-0.993)	0.984(0.974-0.993)	0.998(0.997-1.000)
Flower Chafer							
MV-MNFF (Our Proposal 1)	Full	0.996(0.991-0.999)	0.021(0.019-0.023)	0.996(0.991-0.999)	0.996(0.992-0.999)	0.996(0.991-0.999)	1.000(1.000-1.000)
	w/o CBAM	0.990(0.982-0.997)	0.035(0.032-0.038)	0.990(0.982-0.997)	0.990(0.982-0.997)	0.990(0.982-0.996)	1.000(1.000-1.000)
	w/o MNFF	0.993(0.987-0.998)	0.033(0.031-0.036)	0.992(0.987-0.998)	0.993(0.987-0.998)	0.992(0.987-0.997)	1.000(1.000-1.000)
	w/o CBAM & MNFF	0.995(0.989-1.000)	0.015(0.013-0.017)	0.995(0.989-1.000)	0.995(0.989-0.999)	0.995(0.989-0.999)	1.000(1.000-1.000)
EN-MNFF (Our Proposal 2)	Full	0.995(0.989-0.999)	0.023(0.021-0.025)	0.994(0.988-0.999)	0.994(0.989-0.999)	0.994(0.989-0.998)	1.000(0.999-1.000)
	w/o CBAM	0.997(0.993-0.999)	0.009(0.008-0.010)	0.997(0.992-0.999)	0.996(0.992-0.999)	0.997(0.993-0.999)	1.000(1.000-1.000)
	w/o MNFF	0.998(0.993-1.000)	0.011(0.010-0.012)	0.998(0.993-1.000)	0.998(0.993-1.000)	0.997(0.993-1.000)	1.000(1.000-1.000)
	w/o CBAM & MNFF	0.996(0.992-1.000)	0.011(0.010-0.013)	0.997(0.992-1.000)	0.996(0.992-1.000)	0.997(0.991-1.000)	1.000(1.000-1.000)

Table 5
Performance comparison across fusion settings (3-fold average performance).

Fusion Setting	BA↑ (95% CI)	Loss↓ (95% CI)	Recall↑ (95% CI)	Precision↑ (95% CI)	F1 Score↑ (95% CI)	AUROC↑ (95% CI)
Mealworm						
Both ×	0.990(0.982-0.997)	0.029(0.027-0.031)	0.990(0.983-0.997)	0.990(0.982-0.997)	0.990(0.981-0.997)	1.000(0.999-1.000)
Upper +	0.984(0.975-0.993)	0.040(0.038-0.044)	0.985(0.975-0.993)	0.985(0.975-0.993)	0.984(0.975-0.992)	0.999(0.998-1.000)
Lower +	0.986(0.976-0.994)	0.040(0.037-0.043)	0.985(0.976-0.994)	0.986(0.977-0.994)	0.985(0.976-0.993)	0.999(0.998-1.000)
Both +	0.983(0.972-0.991)	0.052(0.048-0.056)	0.983(0.972-0.992)	0.984(0.973-0.992)	0.983(0.972-0.991)	1.000(0.999-1.000)
Flower Chafer						
Both ×	0.993(0.986-0.998)	0.024(0.021-0.027)	0.992(0.985-0.998)	0.992(0.985-0.998)	0.992(0.985-0.998)	1.000(0.999-1.000)
Upper +	0.993(0.986-0.998)	0.025(0.023-0.028)	0.993(0.987-0.998)	0.993(0.987-0.998)	0.993(0.986-0.998)	1.000(0.999-1.000)
Lower +	0.996(0.992-0.999)	0.016(0.015-0.018)	0.996(0.992-0.999)	0.996(0.992-0.999)	0.996(0.992-0.999)	1.000(1.000-1.000)
Both +	0.992(0.986-0.998)	0.025(0.023-0.028)	0.993(0.986-0.998)	0.993(0.986-0.998)	0.992(0.985-0.998)	1.000(0.999-1.000)
Average						
Both ×	0.992(0.984-0.998)	0.027(0.024-0.029)	0.991(0.984-0.998)	0.991(0.984-0.998)	0.991(0.983-0.998)	1.000(0.999-1.000)
Upper +	0.988(0.980-0.996)	0.033(0.030-0.036)	0.989(0.981-0.996)	0.989(0.981-0.996)	0.988(0.980-0.995)	1.000(0.998-1.000)
Lower +	0.991(0.984-0.996)	0.028(0.026-0.030)	0.990(0.984-0.996)	0.991(0.984-0.996)	0.990(0.984-0.996)	1.000(0.999-1.000)
Both +	0.988(0.979-0.994)	0.038(0.036-0.042)	0.988(0.979-0.995)	0.988(0.980-0.995)	0.988(0.978-0.994)	1.000(0.999-1.000)

Table 6
Comparison with a prior-work baseline versus the proposed MV-MNFF framework (3-fold average performance).

Model	Balanced Accuracy↑ (95% CI)	Loss↓ (95% CI)	Recall↑ (95% CI)	Precision↑ (95% CI)	F1 Score↑ (95% CI)	AUROC↑ (95% CI)
Mealworm						
ResNet-18	0.984(0.974-0.992)	0.068(0.066-0.070)	0.984(0.974-0.992)	0.984(0.975-0.993)	0.984(0.974-0.992)	0.999(0.997-1.000)
MV-MNFF (Our Proposal 1)	0.989(0.982-0.997)	0.032(0.030-0.035)	0.989(0.981-0.996)	0.990(0.982-0.997)	0.989(0.981-0.996)	1.000(0.999-1.000)
Flower Chafer						
ResNet-18	0.995(0.990-0.999)	0.033(0.032-0.035)	0.995(0.989-0.999)	0.995(0.989-0.999)	0.995(0.989-0.999)	1.000(0.999-1.000)
MV-MNFF (Our Proposal 1)	0.996(0.992-0.999)	0.021(0.019-0.023)	0.996(0.992-0.999)	0.996(0.992-0.999)	0.996(0.992-0.999)	1.000(1.000-1.000)

with both modules consistently achieved higher performance than the single-module variants and the variant without both modules. This suggests that jointly combining feature emphasis and feature fusion can strengthen the learned representations for subtle cues. In contrast, EN-MNFF exhibited backbone- and species-dependent behavior: for mealworm datasets, the configuration with the MNFF module removed achieved the best performance, whereas for flower chafer, the configuration with the CBAM module removed showed the highest performance. Overall, the results confirm that CBAM and MNFF jointly stabilize performance in MV-MNFF, whereas their optimal configuration in EN-MNFF varies with model structure and insect species.

Fusion Operation Analysis: To further validate the effectiveness of element-wise multiplication, we compared addition-based variants for the two fusion stages in the module among the three multiplication operations, excluding the channel attention multiplication, which corresponds to channel-wise reweighting rather than feature fusion. Specifically, for DSC with MN in Eq. (2) and MS with MN in Eq. (5), we constructed Upper + and Lower + by replacing each stage with addition, respectively, and Both + by replacing both stages with addition, and compared them with the original Both × setting. Table 5 shows the average classification performance under different fusion settings. For mealworm, the original Both × configuration achieved the best overall performance across the evaluation metrics. In contrast, the addition-based variants generally degraded performance, with the largest drop observed for Both +. For flower chafer, Lower + achieved the best performance on some major metrics, while the original Both × configuration still showed the second-best overall performance. In contrast, Upper + and Both + did not show clear improvements over the original Both ×. Overall, Both × achieved the best overall average performance. These results suggest that element-wise multiplication-based fusion functions as a more effective feature integration mechanism.

Comparison with Prior-Work Baseline: As shown in Table 6, the proposed MV-MNFF model outperformed the prior-work baseline model, ResNet-18, on both mealworm and flower chafer. In particu-

lar, MV-MNFF achieved lower loss and showed consistent improvements across most classification metrics. These results suggest that the MNFF framework more effectively reflects subtle morphological differences.

4.3. Robustness analysis

Stability Analysis: Table 7 presents the stability evaluation of the proposed model under varying Normal-to-Abnormal class ratios for both mealworm and flower chafer datasets. Only the averaged performance across folds is reported in Table 7, while detailed per-fold results are provided in Supplementary Table 4. For mealworm, the model maintained consistently high performance across all class distributions, with average balanced accuracy (BA) ranging from 0.980 to 0.986. AUROC remained above 0.999 under all conditions, and AUPR remained above 0.995. As class imbalance increased, precision and F1 score showed slight decreases, with this tendency becoming more noticeable at the 90:10 ratio. Nevertheless, the overall discriminative performance remained stable. The consistently high AUROC and AUPR indicate that separability between the two classes was preserved across different class-ratio settings. A similar trend was observed for flower chafer. Average BA ranged from 0.994 to 0.996, while AUROC and AUPR consistently remained at high levels. Precision and F1 score also showed slight declines as class imbalance increased, with the reduction becoming relatively larger at the 90:10 ratio. These results demonstrate that the proposed model is robust to class distribution shifts and can reliably operate in real farm environments where population ratios may vary over time.

Sensitivity Analysis: Table 8 summarizes the sensitivity analysis under varying noise levels for the mealworm and flower chafer datasets. Only the averaged performance across folds is reported in Table 8, while detailed per-fold results are provided in Supplementary Table 5. For mealworm, balanced accuracy (BA) remained at or above 0.966 from noise levels 0.01 to 0.05, indicating stable performance under low-noise conditions. However, from noise level 0.10 onward, performance

Table 7
Stability analysis of model performance under varying class ratios (3-fold average).

Species	Ratio	BA↑	Loss↓	Recall↑	Precision↑	F1 Score↑	AUROC↑	AUPR↑
Mealworm	10:90	0.981	0.049	0.980	0.998	0.989	0.999	1.000
	30:70	0.984	0.052	0.980	0.995	0.987	0.999	0.999
	50:50	0.980	0.058	0.980	0.981	0.980	0.999	0.999
	70:30	0.980	0.057	0.979	0.960	0.969	0.999	0.999
	90:10	0.986	0.059	0.991	0.864	0.922	0.999	0.995
Flower Chafer	10:90	0.994	0.017	0.997	0.999	0.998	1.000	1.000
	30:70	0.995	0.020	0.997	0.997	0.997	1.000	1.000
	50:50	0.994	0.023	0.997	0.992	0.995	1.000	1.000
	70:30	0.995	0.027	0.998	0.982	0.990	1.000	0.999
	90:10	0.996	0.028	1.000	0.934	0.965	1.000	1.000

Table 8
Sensitivity analysis of model performance under different noise levels (3-fold average).

Species	Noise	BA↑	Loss↓	Recall↑	Precision↑	F1 Score↑	AUROC↑	AUPR↑
Mealworm	0.00	0.989	0.032	0.990	0.989	0.989	1.000	1.000
	0.01	0.989	0.035	0.990	0.989	0.989	1.000	1.000
	0.05	0.966	0.092	0.939	0.993	0.965	0.998	0.998
	0.10	0.782	0.855	0.565	0.999	0.616	0.988	0.986
	0.20	0.591	6.872	0.182	0.667	0.242	0.873	0.849
Flower Chafer	0.00	0.996	0.021	0.998	0.994	0.996	1.000	1.000
	0.01	0.994	0.025	0.998	0.989	0.994	1.000	1.000
	0.05	0.809	1.376	1.000	0.788	0.865	0.998	0.997
	0.10	0.650	7.192	1.000	0.635	0.761	0.753	0.725
	0.20	0.558	16.923	1.000	0.535	0.696	0.666	0.666

Table 9
Quantitative faithfulness evaluation results of XAI methods (3-fold average performance).

Species	Method	Curve-based		Single-step	
		Pos. Pert. AUC↓	Neg. Pert. AUC↑	Avg. Drop%↓	% Incr. in Confidence↑
Mealworm	LIME	47.57	76.00	6.43	39.65
	SHAP	40.17	75.44	49.94	26.00
	Grad-CAM	70.19	65.78	49.97	25.50
	Saliency Map	40.26	75.78	43.68	26.75
Flower Chafer	LIME	46.25	77.41	9.81	15.15
	SHAP	40.00	66.08	50.24	23.40
	Grad-CAM	69.96	66.45	49.7	23.85
	Saliency Map	40.01	69.69	50.42	23.55

dropped markedly, with BA decreasing to 0.782, recall to 0.565, and F1 score to 0.616. According to the per-fold results in Table S5, this degradation was strongly influenced by a sharp decline in fold 2. AUROC and AUPR also tended to decrease as noise increased, although the decline was relatively limited, remaining at 0.873 and 0.849, respectively, even under the highest-noise condition. For flower chafer, the overall pattern of performance degradation was similar to that of mealworm, although the decline across metrics showed some differences. At noise level 0.01, BA remained high at 0.994, but performance deteriorated rapidly as the noise level increased from 0.05 to 0.20. According to Table S5, the declines observed at noise levels 0.05 and 0.10 were also substantially influenced by inter-fold variability. Unlike mealworm, AUROC and AUPR for flower chafer decreased substantially as well, indicating a more pronounced loss of discriminative ability. In addition, the near-constant recall of 1.000 at noise levels of 0.05 and above appears to reflect overprediction of the positive class rather than true robustness, which is consistent with the concurrent declines in precision and BA. Overall, the proposed model maintained stable performance under low levels of input noise, but performance degradation became evident under high-noise conditions, with the pattern differing by species. These results indicate both the robustness and the limitations of the model and suggest that changes in input quality should be considered in real-world deployment.

4.4. Explainability analysis

Qualitative Visual Explanations: Fig. 3 presents the explainability analysis of the proposed MV-MNFF model using Grad-CAM, Saliency Map, LIME, and SHAP. To enhance robustness, the visual explanation results were averaged across the three folds. For normal samples of both mealworm and flower chafer, all XAI methods consistently highlighted the insect body regions, suggesting that the model primarily relied on object-centered morphological features rather than background cues. Grad-CAM and LIME emphasized the insect regions as the main contributing areas, while SHAP and Saliency Map additionally revealed localized pixel-level relevance around the insect body. In contrast, for abnormal samples, Grad-CAM and LIME exhibited more spatially dispersed activations, suggesting that the evidence at the object level was weakened or appeared less consistently. By comparison, SHAP and Saliency Map maintained relatively clearer localization over the insect regions. These differences imply that abnormal conditions may be characterized by subtle or spatially distributed cues, and that such cues are captured differently depending on the XAI method.

Quantitative Faithfulness Evaluation: To quantitatively support the visualization-based observations, Table 9 presents a comparative evaluation of the faithfulness of visual explanations generated by different XAI methods using positive and negative perturbation AUC,

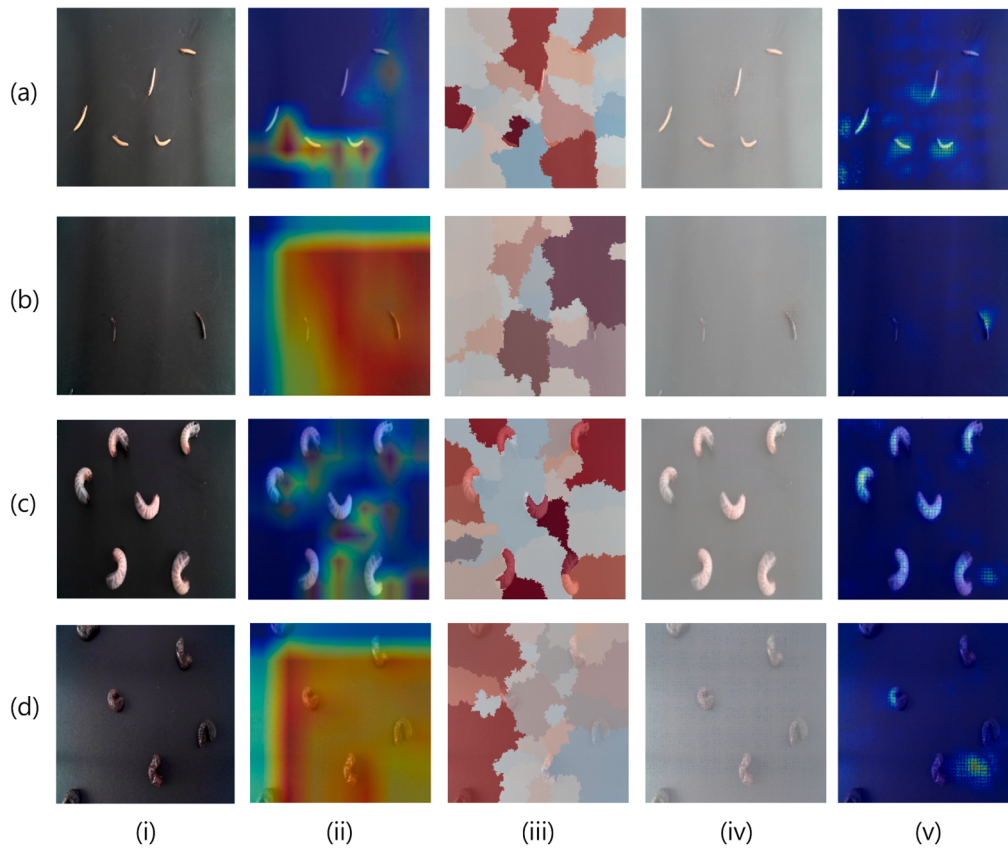


Fig. 3. Explainable AI visualizations of the MV-MNFF. (a) Normal Mealworm, (b) Abnormal Mealworm, (c) Normal Flower Chafer, (d) Abnormal Flower Chafer, (i) Original Image, (ii) Grad-CAM, (iii) LIME (iv) SHAP, (v) Saliency Map.

average drop%, and %increase in confidence. Overall, the patterns of the best-performing methods were similar for mealworm and flower chafer, indicating that the quantitative evaluation results were relatively consistent across species. In the curve-based evaluation, SHAP achieved the lowest positive perturbation AUC for both species, indicating that predictions were most sensitive to the removal of important regions, while LIME achieved the highest negative perturbation AUC, suggesting that predictions remained relatively stable when less important regions were removed. In this regard, Grad-CAM showed noticeably higher positive perturbation AUC and relatively lower negative perturbation AUC for both species, displaying an outlier tendency with more limited faithfulness than the other methods from a curve-based perspective. In the single-step evaluation, SHAP, Grad-CAM, and Saliency Map showed similar levels of average drop% across both species, whereas LIME produced a markedly lower average drop% in both cases, indicating that its highlighted regions alone caused the smallest decrease in prediction confidence. For %increase in confidence, LIME showed the highest value for mealworm, whereas for flower chafer LIME was relatively lower and Grad-CAM showed the highest value, suggesting that confidence gains varied across methods depending on species. Taken together, these quantitative findings indicate an overall consistent pattern of faithfulness, while also revealing metric-dependent differences among XAI methods, with Grad-CAM showing an exceptional tendency in the curve-based evaluation and LIME standing out in the single-step evaluation. Overall, the quantitative faithfulness evaluation in Table 9 supports that the visual patterns observed in Fig. 3 are not merely qualitative impressions but are aligned with changes in model output. In particular, the recurrence of similar best-performing methods across both species strengthens the reproducibility and reliability of the visualization-based explanations. Therefore, the XAI visualization results provide persuasive evidence for the decision basis of MV-MNFF and support its explain-

ability and practical trustworthiness in industrial edible insect quality assessment.

4.5. Web application prototype

Prototype Demonstration: A web-based application prototype was developed to demonstrate the practical feasibility of the proposed model; details and visual examples are provided in Supplementary Fig. 3.

5. Discussion

We proposed a novel Multi-Nonlinear Feature Fusion (MNFF) framework for binary quality classification of edible insects and implemented it as MV-MNFF by integrating the framework with a MambaVision backbone. In this section, we discuss how the proposed MNFF framework differs from existing multi-branch architectures, whether the observed performance gains justify the added architectural complexity, and what limitations and future research directions remain, from both methodological and practical perspectives.

Methodological significance of the MNFF framework: The methodological significance of the MNFF framework lies not simply in adopting a multi-branch structure, but in how post-backbone feature processing is organized within MV-MNFF. After MambaVision extracts hierarchical features, the proposed framework first refines salient representations through CBAM, then strengthens interactions among weak morphological cues through the MNFF module, and finally maps the enhanced representation to a binary decision through a dedicated classification head. In this sense, the proposed framework differs from conventional multi-branch architectures not merely by increasing branch diversity, but by assigning distinct functional roles to successive

components. Within this framework, the core methodological novelty is concentrated in the MNFF module. Unlike conventional approaches that merge parallel branch outputs through simple addition or concatenation, the MNFF module introduces multi-nonlinearity at two successive stages. The first stage applies multi-nonlinearity after depthwise separable convolution, and the second applies it after multi-scale dilated convolution. At both stages, GELU, ELU, and ReLU responses are generated in parallel and fused through multiplicative fusion, enabling the model to emphasize responses that are jointly supported across nonlinear activations and receptive fields. In addition, channel attention recalibrates informative channels, while skip connections preserve low-level spatial information and alleviate optimization instability that may arise from multiplicative fusion. Therefore, the contribution of the MNFF framework lies not in branch proliferation itself, but in an interaction-aware design that progressively combines feature refinement, nonlinear interaction modeling, and decision-oriented representation learning. This interpretation is also consistent with the ablation and fusion analyses, where configurations preserving multi-nonlinearity and multiplicative fusion generally showed stronger and more stable overall performance than simpler alternatives. At the same time, because the magnitude of improvement was not uniform across all settings, the added architectural complexity is most meaningfully interpreted in conditions where weak and spatially dispersed morphological evidence must be explicitly reinforced.

Performance gains and practical implications: The proposed framework showed consistent performance improvements for both insect species, with a larger gain for mealworm and a smaller but stable gain for flower chafer. In addition, MV-MNFF achieved superior overall performance not only relative to the baseline MambaVision but also compared with the prior-work baseline ResNet-18, particularly through lower loss and improved balanced accuracy for both species. Specifically, relative to MambaVision, balanced accuracy improved from 0.970 to 0.989 for mealworm and from 0.992 to 0.996 for flower chafer, while the comparison with ResNet-18 further confirmed that the proposed framework remained competitive against a strong prior benchmark. Robustness analyses also showed that MV-MNFF maintained relatively stable performance under class-ratio shifts, with BA values ranging from 0.980 to 0.986 for mealworm and from 0.994 to 0.996 for flower chafer across class ratios from 10:90 to 90:10. In contrast, sensitivity to noise differed by species. For mealworm, performance remained relatively stable at noise levels of 0.01 and 0.05, but declined substantially from 0.10 onward. For flower chafer, degradation became evident at noise level 0.05 and became more pronounced at 0.10 and 0.20. These results suggest that the proposed framework can operate reliably under moderate perturbations, but still shows vulnerability under severe visual corruption, particularly for flower chafer.

The practical value of the proposed framework is further supported by the explainability analysis. Qualitative visualizations showed that model predictions were primarily driven by insect body regions rather than background noise, and this observation was further supported by quantitative faithfulness evaluation. LIME showed the most consistently favorable results for both species in terms of negative perturbation AUC and average drop%, whereas SHAP achieved the best performance in positive perturbation AUC, indicating that different XAI methods provide complementary strengths. Overall, these findings suggest that MV-MNFF offers not only improved predictive performance, but also the practical trustworthiness required for transparent screening and field-level decision support in industrial-scale edible insect farming.

Limitations: Several limitations should be acknowledged. First, this study was limited to two edible insect species under a binary normal-abnormal classification setting, and thus did not directly establish generalizability to a wider range of species or more fine-grained abnormality categories. Second, although MV-MNFF consistently outperformed the baseline MambaVision and the prior-work baseline ResNet-18, the ablation results suggest that the effectiveness of the proposed modules varies depending on backbone architecture and species characteristics. In par-

ticular, the comparable performance of some reduced configurations indicates that the added architectural complexity of the framework does not necessarily translate into equally large performance gains under all conditions. Third, the robustness analyses were conducted using synthetic noise perturbations and controlled class-ratio shifts, which may not fully capture the complexity of real farm environments. Fourth, although explainability was quantitatively evaluated using multiple faithfulness metrics, these measures remain indirect indicators and therefore cannot fully guarantee that the generated explanations correspond to the true causal basis of model decisions. Finally, while the web-based prototype demonstrated practical feasibility, its real-time operational efficiency and long-term stability in actual farm environments were not systematically validated.

Future directions: Future work should extend the proposed framework to a wider range of edible insect species and more fine-grained abnormality categories beyond the current binary setting. External validation under diverse farm environments, imaging conditions, and breeding densities will also be necessary to assess real-world generalizability. For practical deployment, computational efficiency, lightweight model design, and real-time inference should be further investigated. In addition, integrating uncertainty-aware prediction and temporal or object-level analysis may further improve the reliability of early abnormality screening and decision support in industrial-scale insect farming.

6. Conclusion

We presented two complementary approaches for edible insect quality classification: a flexible Multi-Nonlinear Feature Fusion (MNFF) framework and its instantiation as a binary classifier, MV-MNFF, based on the MambaVision backbone. The MNFF framework integrates CBAM, a task-specific classification head, and a novel MNFF module that combines depthwise separable convolution with multi-nonlinearity, multi-scale dilated convolution with multi-nonlinearity, channel attention, and skip connection, enabling effective representation of fine-grained morphological characteristics in edible insects.

Experimental evaluations showed that MV-MNFF consistently achieved superior classification performance for both insect species. Stability and sensitivity analyses further indicated that the proposed model remained relatively stable under class-ratio shifts and low-to-moderate noise conditions, while substantial degradation emerged under high-noise settings, particularly for flower chafer. From an explainability perspective, multiple XAI methods and quantitative faithfulness-based evaluation verified that the model's predictions were primarily driven by insect body regions rather than background artifacts. In addition, a web application prototype demonstrated the potential practical utility of the proposed approach in farm environments.

Overall, the proposed MNFF-based framework provides a robust, explainable, and practically deployable solution for data-driven edible insect quality management in agricultural settings.

Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

Competing interests

The authors declare that there is no conflict of interest in this paper.

Authors' contributions

S.-H. Choi contributed to the research concept and design. Data collection was performed by Hye-young Wang. Data analysis and interpretation were carried out by Hye-young Wang and S.-H. Choi. Hye-young Wang and S.-H. Choi drafted the manuscript and critically revised it. The final approval of the version to be published was given by S.-H. Choi.

CRediT authorship contribution statement

Hye-Young Wang: Writing – original draft, Software, Conceptualization; **Seoung-Ho Choi:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Conceptualization.

Data availability

The data supporting the findings of this study are available from the public data portal. <https://www.aihub.or.kr/aihubdata/data/view.do?srchOptnCnd=OPTNCND001&currMenu=115&topMenu=100&searchKeyword=%EA%B3%A4%EC%B6%A9&aihubDataSe=data&dataSetSn=71701>

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by [Hansung University](#).

Supplementary material

Supplementary material associated with this article can be found in the online version at [10.1016/j.rineng.2026.110313](https://doi.org/10.1016/j.rineng.2026.110313).

References

- [1] S. Belluco, A. Halloran, A. Ricci, New protein sources and food legislation: the case of edible insects and EU law, *Food Secur.* 9 (4) (2017) 803–814. <https://doi.org/10.1007/s12571-017-0704-0>
- [2] A. van Huis, J.K. Tomberlin (Eds.), *Insects as Food and Feed: From Production to Consumption*, Series of Insects as Food and Feed, Wageningen Academic Publishers, Wageningen, The Netherlands, 2017.
- [3] B.A. Rumpold, O.K. Schlüter, Nutritional composition and safety aspects of edible insects, *Mol. Nutr. Food Res.* 57 (5) (2013) 802–823. <https://doi.org/10.1002/mnfr.201200735>
- [4] S. Belluco, C. Losasso, M. Maggioletti, C.C. Alonzi, M.G. Paoletti, A. Ricci, Edible insects in a food safety and nutritional perspective: a critical review, *Compr. Rev. Food Sci. Food Saf.* 12 (3) (2013) 296–313. <https://doi.org/10.1111/1541-4337.12014>
- [5] J. Eilenberg, J.M. Vlak, C. Nielsen-LeRoux, S. Cappellozza, A.B. Jensen, Diseases in insects produced for food and feed, *J. Insects Food Feed* 1 (2) (2015) 87–102. <https://doi.org/10.3920/JIFF2014.0022>
- [6] M.C. Ogwu, M. Osei-Kwarteng, F.E. Mukah, Insect farming: methods and sustainability, in: M. Chidozie Ogwu, S. Chibueze Izah (Eds.), *Edible Insects: Nutritional Benefits, Culinary Innovations and Sustainability*, Springer, Cham, 2025, pp. 69–95. https://doi.org/10.1007/978-3-031-90087-7_4
- [7] P. Majewski, P. Zapotoczny, P. Lampa, R. Burduk, J. Reiner, Multipurpose monitoring system for edible insect breeding based on machine learning, *Sci. Rep.* 12 (2022). <https://doi.org/10.1038/s41598-022-11794-5>
- [8] F. Ssemakula, S. Nawoya, C.N. Kunyanga, R. Akol, D. Nakimbugwe, R.C. Bett, H. Karstoft, K. Bjerger, A. Katumba, C. Mwikirize, G. Gebreyesus, Emerging technologies for fast determination of nutritional quality and safety of insects for food and feed: a review, *Comput. Electron. Agric.* 239 (2025) 111126. <https://doi.org/10.1016/j.compag.2025.111126>
- [9] A. Halloran, R. Flore, P. Vantomme, N. Roos (Eds.), *Edible Insects in Sustainable Food Systems*, Springer International Publishing, Cham, Switzerland, 2018. <https://doi.org/10.1007/978-3-319-74011-9>
- [10] J. Muinde, C.M. Tanga, J. Olukuru, C. Odhiambo, H.E.Z. Tonnang, K. Senagi, Application of machine learning techniques to discern optimal rearing conditions for improved black soldier fly farming, *Insects* 14 (5) (2023) 479. <https://doi.org/10.3390/insects14050479>
- [11] S. Nawoya, F. Ssemakula, R. Akol, Q. Geissmann, H. Karstoft, K. Bjerger, C. Mwikirize, A. Katumba, G. Gebreyesus, Computer vision and deep learning in insects for food and feed production: a review, *Comput. Electron. Agric.* 216 (2024) 108503. <https://doi.org/10.1016/j.compag.2023.108503>
- [12] M.F. Hansen, A. Oparaake, R. Gallagher, A. Karimi, F. Tariq, M.L. Smith, Towards machine vision for insect welfare monitoring and behavioural insights, *Front. Vet. Sci.* 9 (2022) 835529. <https://doi.org/10.3389/fvets.2022.835529>
- [13] N. Giulietti, P. Castellini, C. Truzzi, B. Ajdini, M. Martarelli, Vision measurement system for gender-based counting of *acheta domesticus*, *Sensors* 24 (15) (2024) 4936. <https://doi.org/10.3390/s24154936>
- [14] A. Lee, G. Kim, S.-J. Hong, S.-W. Kim, G. Kim, Classification of dead cocoons using convolutional neural networks and machine learning methods, *IEEE Access* 11 (2023) 137317–137327. <https://doi.org/10.1109/ACCESS.2023.3338540>
- [15] X.-S. Wei, Y.-Z. Song, O.M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: a survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (12) (2022) 8927–8948. <https://doi.org/10.1109/TPAMI.2021.3126648>
- [16] G. Yang, Y. He, Y. Yang, B. Xu, Fine-grained image classification for crop disease based on attention mechanism, *Front. Plant Sci.* Volume 11 - 2020 (2020). <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2020.600854>
- [17] B. Guan, Y. Wu, J. Zhu, J. Kong, W. Dong, GC-faster RCNN: The object detection algorithm for agricultural pests based on improved hybrid attention mechanism, *Plants* 14 (7) (2025). <https://doi.org/10.3390/plants14071106>
- [18] Z. Zhang, Z. Lei, M. Omura, H. Hasegawa, S. Gao, Dendritic learning-incorporated vision transformer for image recognition, *IEEE/CAA J. Autom. Sin.* 11 (2) (2024) 539–541. <https://doi.org/10.1109/JAS.2023.123978>
- [19] A. Hatamizadeh, J. Kautz, Mambavision: a hybrid mamba-transformer vision backbone, in: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 25261–25270.
- [20] J. An, Y. Du, P. Hong, L. Zhang, X. Weng, Insect recognition based on complementary features from multiple views, *Sci. Rep.* 13 (2023). <https://doi.org/10.1038/s41598-023-29600-1>
- [21] M. Prunella, R.M. Scardigno, D. Buongiorno, A. Brunetti, N. Longo, R. Carli, M. Dotoli, V. Bevilacqua, Deep learning for automatic vision-based recognition of industrial surface defects: a survey, *IEEE Access* 11 (2023) 43370–43423. <https://doi.org/10.1109/ACCESS.2023.3271748>
- [22] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, Learning deep bilinear transformation for fine-grained image representation, in: H. Wallach, H. Larochelle, A. Beygelzimer, F.D. Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, 32, Curran Associates, Inc., 2019, pp. 4279–4288. https://proceedings.neurips.cc/paper_files/paper/2019/file/959ef477884b6ac2241b19ee4fb776ae-Paper.pdf
- [23] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1449–1457.
- [24] G. Sun, H. Cholakkal, S. Khan, F. Khan, L. Shao, Fine-grained recognition: accounting for subtle differences between similar classes, *Proc. AAAI Conf. Artif. Intell.* 34 (07) (2020) 12047–12054. <https://doi.org/10.1609/aaai.v34i07.6882>
- [25] T.T. Høye, J. Årje, K. Bjerger, O.L.P. Hansen, A. Iosifidis, F. Leese, H.M.R. Mann, K. Meissner, C. Melvad, J. Raitoharju, Deep learning and computer vision will transform entomology, *Proc. Natl. Acad. Sci.* 118 (2) (2021) e2002545117. <https://doi.org/10.1073/pnas.2002545117>
- [26] A.C. Teixeira, J. Ribeiro, R. Morais, J.J. Sousa, A. Cunha, A systematic review on automatic insect detection using deep learning, *Agriculture* 13 (3) (2023). <https://doi.org/10.3390/agriculture13030713>
- [27] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, J. Yang, IP102: a large-scale benchmark dataset for insect pest recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8787–8796.
- [28] F. Ren, W. Liu, G. Wu, Feature reuse residual networks for insect pest recognition, *IEEE Access* 7 (2019) 122758–122768. <https://doi.org/10.1109/ACCESS.2019.2938194>
- [29] W. Liu, G. Wu, F. Ren, Deep multibranch fusion residual network for insect pest recognition, *IEEE Trans. Cognit. Dev. Syst.* 13 (3) (2021) 705–716. <https://doi.org/10.1109/TCDS.2020.2993060>
- [30] A. Baur, D. Koch, B. Gattermig, A. Delgado, Noninvasive monitoring system for Tenebrio molitor larvae based on image processing with a watershed algorithm and a neural net approach, *J. Insects Food Feed* 8 (8) (2022) 913–920. <https://doi.org/10.3920/JIFF2021.0185>
- [31] A. Pratondo, A. Bramantoro, Classification of Zophobas morio and Tenebrio molitor using transfer learning, *PeerJ Comput. Sci.* 8 (2022) e884. <https://doi.org/10.7717/peerj-cs.884>
- [32] K. Mungase, S. Chiwhane, P. Paygude, CNN-based image classification of silkworm for early prediction of diseases, *Comput. Sci. Math. Forum* 12 (1) (2025) 14. <https://doi.org/10.3390/cmsf2025012014>
- [33] Y. Rao, G. Chen, J. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1025–1034.
- [34] T. Zheng, X. Yang, J. Lv, M. Li, S. Wang, W. Li, An efficient mobile model for insect image classification in the field pest management, *Eng. Sci. Technol. Int. J.* 39 (2023) 101335. <https://doi.org/10.1016/j.jestech.2023.101335>
- [35] H. He, S. Zhu, L. Shen, X. Chang, Y. Wang, D. Zeng, B. Xiong, F. Dai, T. Zhao, Integrated analysis of machine learning and deep learning in silkworm pupae (bombyx mori) species and sex identification, *Animals* 13 (23) (2023). <https://doi.org/10.3390/ani13233612>
- [36] P. Majewski, P. Lampa, R. Burduk, J. Reiner, End-to-end solution for tenebrio molitor rearing monitoring with uncertainty estimation and domain shift detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2024, pp. 5498–5507. <https://doi.org/10.1109/CVPRW63382.2024.00559>
- [37] Y. Lin, J. Ma, Q. Wang, D.-W. Sun, Applications of machine learning techniques for enhancing nondestructive food quality and safety detection, *Crit. Rev. Food Sci. Nutr.* 63 (12) (2023) 1649–1669. PMID: 36222697, <https://doi.org/10.1080/10408398.2022.2131725>
- [38] C. Yu, X. Zhao, Q. Zheng, P. Zhang, X. You, Hierarchical bilinear pooling for fine-grained visual recognition, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 574–589.

- [39] M.F. Hansen, A. Oparaeke, R. Gallagher, A. Karimi, F. Tariq, M.L. Smith, Towards machine vision for insect welfare monitoring and behavioural insights, *Front. Vet. Sci.* Volume 9 - 2022 (2022). <https://doi.org/10.3389/fvets.2022.835529>
- [40] P. Liu, X. He, K. Zhao, W. Li, B. Huang, Physiological state recognition model of small silkworm based on improved YOLOv5, *Trans. Inst. Meas. Control* (2024). First published online: 13 Nov 2024, <https://doi.org/10.1177/00368504241298136>
- [41] H. Shi, W. Xiao, S. Zhu, L. Li, J. Zhang, CA-YOLOv5: detection model for healthy and diseased silkworms in mixed conditions based on improved YOLOv5, *Int. J. Agric. Biol. Eng.* 16 (6) (2023). <https://doi.org/10.25165/j.ijabe.20231606.7854>
- [42] D.R. Bourne, C.J. Kyle, H.N. LeBlanc, D. Beresford, Technical note: a rapid, non-invasive method for measuring live or preserved insect specimens using digital image analysis, *Forensic Sci. Int. Synergy* 1 (2019) 140–145. <https://doi.org/10.1016/j.fsisyn.2019.07.006>
- [43] S. Kong, C. Fowlkes, Low-rank bilinear pooling for fine-grained classification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 365–374.
- [44] L. Chi, Z. Yuan, Y. Mu, C. Wang, Non-local neural networks with grouped bilinear attentional transforms, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11804–11813.
- [45] M. Tan, Q. Le, EfficientNetV2: smaller models and faster training, in: *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. <https://proceedings.mlr.press/v139/tan21a.html>.
- [46] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, X. Pan, Next-ViT: next generation vision transformer for efficient deployment in realistic industrial scenarios, 2022, [arXiv:2207.05501](https://arxiv.org/abs/2207.05501),
- [47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [48] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 2261–2269. <https://doi.org/10.1109/CVPR.2017.243>
- [49] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM + + : generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847. <https://doi.org/10.1109/WACV.2018.00097>
- [50] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 782–791.