BMC Biotechnology

RESEARCH

Open Access

Explainable deep learning ensemble framework for accurate classification of wild poisonous mushroom species

Aras Fahrettin Korkmaz[1], Fatih Ekinci[2], Eda Kumru[3], Abdullah Aydoğan[4], Hasna Sena Kaymak[4], Mustafa Sevindik[5*], Mehmet Serdar Güzel[4] and Ilgaz Akata[6]

Abstract

This study introduces an explainable deep learning framework for the accurate classification of wild poisonous mushroom species, contributing to food safety. A balanced dataset of 3600 high-resolution images representing 18 species was employed, split into training, validation, and test subsets. To enhance variability and reduce overfitting, the images were expanded using advanced augmentation techniques (rotation, flipping, brightness/contrast adjustments, noise injection, etc.), increasing the training set to 7200 samples. Four pretrained CNN architectures DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2 were fine-tuned via transfer learning and evaluated with multiple performance metrics. Among the individual models, EfficientNet-B3 achieved the highest accuracy of 93.0%. However, ensemble strategies based on soft voting consistently outperformed single models. The four-model ensemble (DenseNet121, EfficientNet-B3, MobileNet-V3, ShuffleNet-V2) achieved the best results with 95.67% accuracy, 95.42% MCC, and a log loss of 0.175. Explainable AI methods (Grad-CAM, Grad-CAM++) revealed that classification decisions corresponded to biologically meaningful regions, thereby improving interpretability and reliability. This study holds direct life-saving potential by reducing poisoning incidents caused by misidentifications. In addition, it contributes to food safety by supporting reliable identification of toxic species within the agricultural and food supply chain. Furthermore, it pioneers the integration of AI methodologies in fungal taxonomy, providing a robust foundation for future ecological, agricultural, and biotechnological research.

Keywords Deep learning, Poisonous mushroom, Mushroom classification, Ensemble models, Explainable AI, Computer vision

*Correspondence:
Mustafa Sevindik
sevindik27@gmail.com
[1]Faculty of Health Sciences, Dietetics and Nutrition Department, İstanbul Kültür University, Şirinevler Campus, Istanbul 34191, Türkiye
[2]Institute of Artificial Intelligence, Ankara University, Ankara 06100, Türkiye
[3]Graduate School of Natural and Applied Sciences, Ankara University, Ankara 06830, Türkiye
[4]Department of Computer Engineering, Faculty of Engineering, Ankara University, Ankara 06830, Türkiye
[5]Department of Biology, Faculty of Engineering and Natural Sciences, Osmaniye Korkut Ata University, Ankara, Türkiye
[6]Department of Biology, Faculty of Science, Ankara University, Ankara 06100, Türkiye

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Globally, nearly 150,000 mushroom species have been identified, but only about 2000 are edible, with roughly 700 recognized for medicinal uses. Many species are also toxic, leading to fatal accidents each year, often caused by misidentification [1]. Poisonous mushrooms contain bioactive compounds such as small organic molecules, peptides, and proteins, which can cause effects ranging from mild gastrointestinal discomfort to severe liver and kidney damage, sometimes resulting in death [2]. Most mushroom poisonings are linked to a few primary toxins: amatoxins, phallotoxins, orellanine, muscarine, psilocybin, illudin, and coprine [3].

Amanita phalloides, *A. virosa,* and *Galerina marginata* contain amatoxins, a group of deadly toxins [4, 5]. *Amanita muscaria* and *A. pantherina* produce isoxazole compounds, distinctive metabolites of these mushrooms [6]. *Gyromitra esculenta* contains Gyromitrin, a toxic compound known for its unique structure and effects on health [7]. *Coprinopsis atramentaria* contains Coprine, an alcohol dehydrogenase inhibitor that causes the alcohol flush reaction [8]. *Inocybe geophylla*, *Mycena pura*, and *M. rosea* contain muscarine, an alkaloid traditionally linked to mushroom poisoning [6]. *Gymnopilus junonius* is a toxic, hallucinogenic mushroom associated with unusual behavior and uncontrollable laughter [9]. *Paxillus involutus* is deadly, causing rare but fatal immune-mediated hemolysis with rapid gastrointestinal and cardiovascular symptoms [10]. *Agaricus xanthodermus, Hygrophoropsis aurantiaca, Hypholoma fasciculare, Rubroboletus satanas,* and *Tricholoma sulphureum* can cause gastrointestinal poisoning, but the specific toxins responsible are still unknown [3].

Accurately identifying poisonous mushrooms is difficult because many toxic species resemble edible ones and can only be distinguished by subtle traits. This often leads to misidentification with serious consequences [11]. Traditional taxonomy is slow, requires expertise, and visual similarities can obscure correct identification [12]. AI offers a solution, as machine learning and computer vision can effectively recognize and differentiate fungal species [13–18]. Incorporating AI could greatly improve the accuracy of poisonous mushroom identification.

In deep learning-based image classification tasks, the combined use of diverse architectures and auxiliary techniques plays a vital role in improving performance and producing generalizable results [19]. In this framework, convolutional neural network (CNN) architectures such as DenseNet121, EfficientNet-B3 [20], MobileNet-V3 [21], and ShuffleNet-V2 [22], pretrained on ImageNet, were fine-tuned through transfer learning [23]. To boost data diversity and reduce overfitting, advanced augmentation techniques (e.g., rotations, flips, brightness/contrast adjustments, Gaussian noise, and scaling) were applied using the Albumentations library [24]. To further enhance classification accuracy, ensemble strategies based on soft voting were employed, with model weights optimized via an exhaustive grid-search procedure [25]. Additionally, explainable artificial intelligence (XAI) methods such as Grad-CAM and Grad-CAM++ were integrated to ensure that classification decisions aligned with biologically meaningful regions [26, 27]. The models were thoroughly evaluated using multidimensional metrics, including Accuracy, Precision, Recall, F1-Score, Log Loss, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC) [28, 29]. In recent years, machine learning and deep learning methods have been increasingly used for mushroom classification, especially for toxic species.

The primary objective of this study is to develop a reliable and interpretable AI-based framework for the classification of poisonous mushroom species. Beyond improving classification accuracy, this work makes a distinct contribution to the literature by jointly addressing performance, robustness, and explainability—aspects that are often treated separately in previous studies. Unlike earlier approaches that rely on limited datasets or single-model evaluations, the proposed methodology employs a fully balanced multi-class dataset and integrates multiple state-of-the-art CNN architectures through a weighted ensemble strategy. Moreover, the systematic incorporation of explainable AI techniques enables biologically meaningful interpretation of model decisions, directly linking learned features to key morphological traits. By unifying high accuracy with transparent decision-making, this study fills an important methodological gap and establishes a transferable reference framework for future research in fungal taxonomy, food safety, agricultural monitoring, and AI-driven biological classification.

Materials and methods

The methodological workflow followed in this study is as follows: (i) First, a balanced dataset was constructed comprising 18 poisonous mushroom species, with 200 images per class; the majority of the images were obtained from verified open-access sources such as GBIF, and species identification was visually revalidated by an expert mycologist. (ii) The dataset was split into training, validation, and test sets in a 50:25:25 ratio while preserving class balance. (iii) Data augmentation was applied exclusively to the training set to reduce overfitting and increase variability, using the Albumentations library with transformations including rotation, horizontal and vertical flipping, brightness and contrast adjustment, Gaussian noise, blurring, and shift–scale–rotate operations, thereby expanding the number of training samples. (iv) Subsequently, four ImageNet-pretrained

CNN architectures—DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2—were fine-tuned via transfer learning; the final classification layer of each architecture was adapted to 18 classes, and inputs were resized to architecture-specific resolutions and normalized using ImageNet statistics. (v) The models were trained for a fixed number of epochs using the Adam optimizer and categorical cross-entropy loss; the model weights achieving the best validation performance were saved and evaluated on the test set. (vi) Following single-model evaluations, ensemble structures consisting of pairwise, triple, and four-model combinations were constructed; for each ensemble, a weighted soft-voting strategy was employed, with weights selected via a 0.1-step exhaustive grid search on the validation set. (vii) Performance was assessed for both individual models and ensembles using multiple metrics, including Accuracy, Precision, Recall, F1-score, AUC, MCC, and Log Loss. (viii) Finally, to examine whether the decision-making process relied on biologically meaningful regions, explainability analyses were conducted using Grad-CAM and Grad-CAM++; ensemble explanations were obtained by combining the heatmaps of the constituent models according to their selected weights.

The images used in this study were primarily obtained from the Global Biodiversity Information Facility (GBIF, [30]), which provides publicly available and taxonomically verified records curated by professional mycologists. Each image in the dataset corresponds to an officially identified specimen within the GBIF repository. Although no new physical specimens were collected or deposited in a herbarium, the taxonomic validity of all mushroom species used in this study was visually re-verified by Prof. Dr. Ilgaz Akata (Department of Biology, Faculty of Science, Ankara University, Türkiye), an expert mycologist specializing in fungal taxonomy. As the study relied exclusively on digital image data from verified public databases, no physical voucher specimens were created or deposited, but GBIF accession references are available upon request.

The dataset used in this study consists of high-resolution images of 18 different fungi species: *Agaricus xanthodermus, Amanita muscaria, Amanita pantherina, Amanita phalloides, Amanita virosa, Coprinopsis atramentaria, Galerina marginata, Gymnopilus junonius, Gyromitra esculenta, Hygrophoropsis aurantiaca, Hypholoma fasciculare, Inocybe geophylla, Mycena pura, Mycena rosea, Omphalotus olearius, Paxillus involutus, Rubroboletus satanas*, and *Tricholoma sulphureum* (Figure 1a, 1b, 1c).

A total of 3600 images were collected, with exactly 200 samples allocated per class, ensuring perfect class balance. Approximately 5% of the images were captured directly by the authors using high-resolution cameras. In comparison, the remaining 95% were obtained from publicly accessible biodiversity repositories, primarily the Global Biodiversity Information Facility (GBIF) https://www.gbif.org [30]. All images were stored in JPEG format with a consistent resolution of 300 dpi. Given the uniform distribution across all classes, the dataset is considered highly suitable for multi-class classification tasks. A summary of the image format, resolution, and source distribution per species is presented in Table 1.

The dataset, consisting of 3600 original images from 18 balanced classes, was partitioned into training, validation, and test sets in a 50:25:25 ratio, corresponding to 1800, 900, and 900 images, respectively, with class balance preserved across all splits [31]. Data augmentation was applied exclusively to the training set to increase variability and reduce overfitting, whereby each original image was transformed into three additional variants using the Albumentations library, yielding 5400 augmented samples and expanding the training set to 7200 photos. The augmentation techniques included horizontal and vertical flips, 90° rotations, brightness/contrast adjustments, Gaussian noise, blurring, and random shift–scale–rotate operations. At the same time, all images were normalized using the ImageNet mean and standard deviation before model training [31].

The classification framework in this study leverages four pretrained convolutional neural network architectures: DenseNet121 [32], EfficientNet-B3 [33], MobileNet-V3 [34], and ShuffleNet-V2 [35], each initialized with ImageNet weights and employed at their recommended input resolutions (224×224 for DenseNet121, MobileNet-V3, and ShuffleNet-V2; 300×300 for EfficientNet-B3). DenseNet121 adopts dense connectivity to promote feature reuse and mitigate the vanishing gradient problem, enabling efficient information flow across layers [32]. EfficientNet-B3 utilizes compound scaling to balance network depth, width, and resolution, achieving a substantial trade-off between accuracy and computational cost [20]. MobileNet-V3, designed through neural architecture search and optimized for speed, is well-suited for resource-constrained environments while maintaining competitive performance [36]. ShuffleNet-V2 incorporates a channel shuffle mechanism to enhance efficiency by reducing computational complexity without significantly sacrificing accuracy [22]. The diversity in architectural design, input resolution, and computational requirements among these models provides a robust foundation for comparative performance evaluation in the multi-class mushroom classification task. As shown in Fig. 2, four pretrained CNN architectures are compared in terms of parameter count, computational cost, inference speed, model depth, and representational capacity.

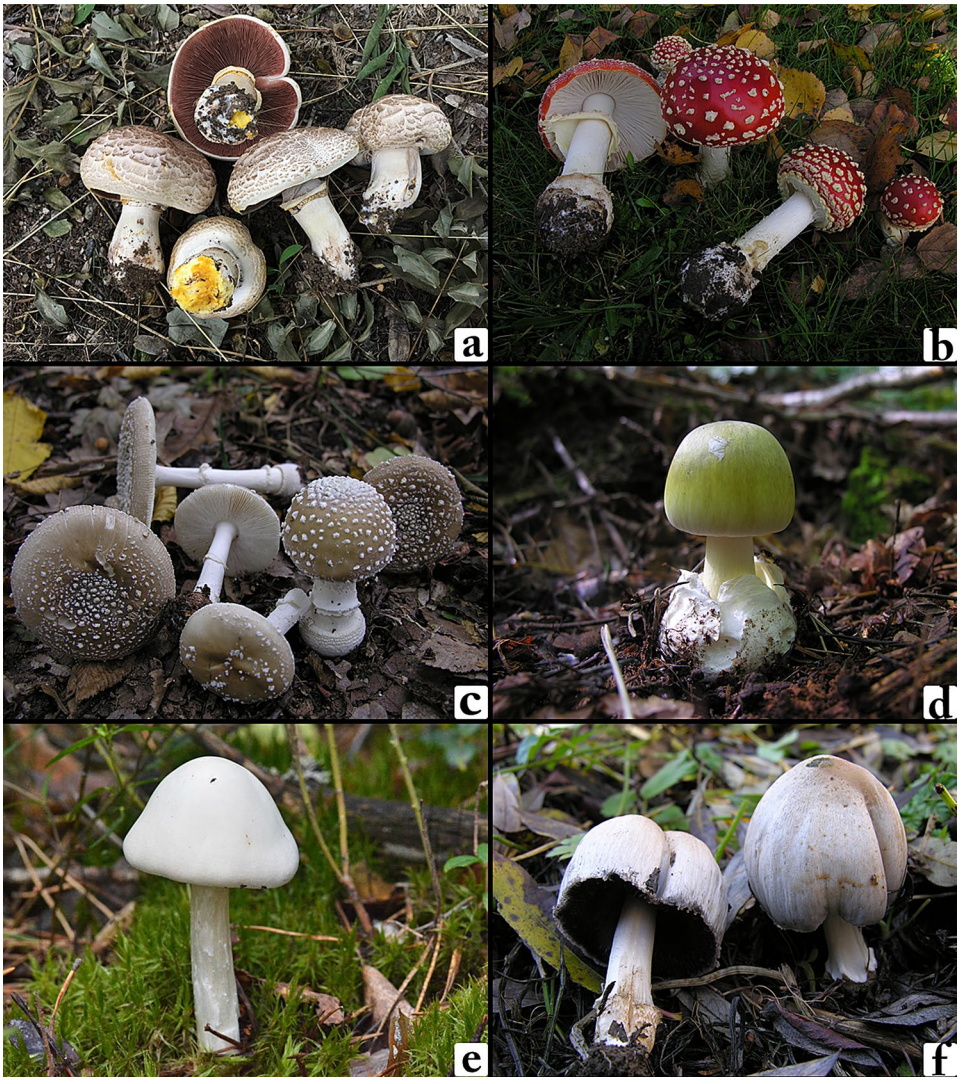

Fig. 1 (**a**) *Agaricus xanthodermus*, (**b**) *Amanita muscaria*, (**c**) *Amanita pantherina*, (**d**) *Amanita phalloides*, (**e**) *Amanita virosa*, (**f**) *Coprinopsis atramentaria*. b. (**a**) *Galerina marginata*, (**b**) *Gymnopilus junonius*, (**c**) *Gyromitra esculenta*, (**d**) *Hygrophoropsis aurantiaca*, (**e**) *Hypholoma fasciculare*, (**f**) *Inocybe geophylla*.c. (**a**) *Mycena pura*, (**b**) *Mycena rosea*, (**c**) *Omphalotus olearius*, (d) *Paxillus involutus*, (**e**) *Rubroboletus satanas*, (**f**) *Tricholoma sulphureum*

Figure 2 presents a comparative profile of four pre-trained CNN architectures DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2 across five key characteristics: parameter count, computational cost (GFLOPs), inference speed, model depth, and representational capacity. All metrics were normalized to a 0–10 scale using data-driven bounds with 10% padding around the observed minimum and maximum, with cost-type metrics (GFLOPs) inverted to ensure that higher values consistently indicate more favorable performance. In terms of parameter count, the models range from 2.28 M (ShuffleNet-V2) to 12.23 M (EfficientNet-B3), with DenseNet121 (7.98 M) and MobileNet-V3 (5.48 M) occupying the mid-range; EfficientNet-B3's larger size supports higher representational capacity, while ShuffleNet-V2's compact design emphasizes efficiency. GFLOPs values span from 0.14 (ShuffleNet-V2) to 2.83 (DenseNet121), with MobileNet-V3 at 0.22 and EfficientNet-B3 at 1.83, indicating that the mobile-oriented models impose lower computational demands and are better suited for resource-constrained environments. For inference speed, MobileNet-V3 (250 FPS) and ShuffleNet-V2 (230 FPS) lead, followed by DenseNet121 (110 FPS) and EfficientNet-B3 (90 FPS), underscoring the real-time suitability of lightweight designs. In model depth, EfficientNet-B3 is the deepest (300 layers), followed by DenseNet121 (121 layers), and then MobileNet-V3 (75 layers) and ShuffleNet-V2 (50 layers), which are shallower, reflecting a trade-off between complexity and efficiency. Finally, in representational capacity measured by Top-1 ImageNet accuracy, EfficientNet-B3 achieves the highest score (82.0%), followed by MobileNet-V3 (75.3%),



Fig. 1 (continued)

DenseNet121 (74.4%), and ShuffleNet-V2 (69.4%), demonstrating that deeper and larger models generally yield stronger discriminative performance, albeit with higher computational costs.

$$L_m = 0.9 \times min_j \{x_{m,j}\} \quad (1)$$

$$U_m = 1.1 \times max_j \{x_{m,j}\} \quad (2)$$

$$\tilde{x}_{m,j} = min\left(max\left(x_{m,j}\ ,\ L_m\right), U_m\right) \quad (3)$$

$$s_{m,j} = 10 \times \left(\tilde{x}_{m,j}\ -\ L_m\right) / \left(U_m -\ L_m\right) \quad (4)$$

$$s^{final}{}_{m,j} = \begin{cases} 10 - s_{m,j}, if\ m\ is\ cost - type\ metrics \\ s_{m,j}, otherwise \end{cases} \quad (5)$$

The variables are defined as follows: $[L_m]$ denotes the lower bound for metric m, computed as 90% of the observed minimum across models (Eqs. 1, 3 and 4); $x_{m,j}$ is the raw (unnormalized) value of metric m for model j (Eqs. 1, 2 and 3); U_m denotes the upper bound for metric m, computed as 110% of the observed maximum across models (Eqs. 2, 3 and 4); $\tilde{x}_{m,j}$ is the clamped value of $x_{m,j}$ restricted to the interval $[L_m\ , U_m]$ (Eq. 3); $s_{m,j}$ is the normalized score obtained by mapping $\tilde{x}_{m,j}$ to the $[0, 10]$ range using L_m and U_m; and $s^{final}{}_{m,j}$ is the post-inversion score for cost-type metrics (e.g., GFLOPs) so that larger values consistently reflect more favorable characteristics (Eqs. 4 and 5). Here, m indexes the five characteristics (parameter count, GFLOPs, inference speed, model depth, and representational capacity), and j indexes the four CNN models (Fig. 3) (Eqs. 1, 2, 3, 4 and 5) [35].

The ensemble module constructs latefusion models by aggregating the softmax probability vectors of the base networks DenseNet121, EfficientNet-B3, MobileNet-V3,



Fig. 1 (continued)

and ShuffleNet-V2 over all pairwise, triple, and four-way combinations (i.e., $\binom{4}{2} = 6$, $\binom{4}{3} = 4$, $\binom{4}{4} = 1$; 11 ensembles in total) [35]. For a given combination, class-posterior probabilities are fused via weighted soft voting at the probability level, $\hat{p} = \sum_k w_k p^{(k)}$ under the simplex constraint $w_k \geq 0$ and $\sum_k w_k = 1$. The weights are selected by an exhaustive grid search with step size 0.1 (all weight tuples summing to 1 are considered), and the best tuple is chosen on the validation set using a lexicographic criterion that prioritizes, in order: accuracy and macroF1 (maximize), logloss (minimize), macro AUCOVR (maximize), macro precision (maximize), macro recall (maximize), and MCC (maximize). For each bestweight ensemble, we report the full metric suite and save the corresponding confusion matrix; probability calibration is not applied, and fusion operates strictly on the modelreported probabilities. This strategy searches both the combinatorics of models and their contribution weights explicit and reproducible while keeping the aggregation rule simple and wellposed [36].

Notation

$p^{(k)}$: The $C-$dimensional softmax probability vector produced by base model k.

w_k: The weight assigned to model k in the ensemble; all w_k satisfy $w_k \geq 0$ and $\sum_k w_k = 1$.

$\hat{p}$: The final ensemble prediction probability vector obtained through weighted probability fusion.

k: Index of the model in the ensemble

C: Number of classes

All models were initialized with pretrained ImageNet weights to leverage transfer learning benefits. The final classification layer of each architecture was replaced with

Table 1 Mushroom species, photograph sources, and approximate continents [30]

Mushroom Species Name	% of photographs received from source	The Continents Of Capture
Agaricus xanthodermus	> 95%	North America, Europe, Australia
Amanita muscaria	> 95%	North America, Europe, Asia
Amanita pantherina	> 95%	North America, Europe, Asia
Amanita phalloides	> 95%	North America, Europe, Asia
Amanita virosa	> 95%	North America, Europe
Coprinopsis atramentaria	> 95%	North America, Europe, Asia, Australia
Galerina marginata	> 95%	North America, Europe, Asia
Gymnopilus junonius	> 95%	North America, Europe, Australia, Africa
Gyromitra esculenta	> 95%	North America, Europe, Asia
Hygrophoropsis aurantiaca	> 95%	North America, Europe, Asia
Hypholoma fasciculare	> 95%	North America, Europe, Asia, Australia
Inocybe geophylla	> 95%	North America, Europe, Asia
Mycena pura	> 95%	North America, Europe, Asia, Australia
Mycena rosea	> 95%	Europe, Asia
Omphalotus olearius	> 95%	North America, Europe
Paxillus involutus	> 95%	North America, Europe, Asia
Rubroboletus satanas	> 95%	Europe
Tricholoma sulphureum	> 95%	North America, Europe, Asia

a fully connected layer matching the number of target classes, enabling task-specific adaptation. Fine-tuning was applied to the entire network instead of restricting training to only the classifier layer, allowing all parameters to be updated during optimization. Model inputs were resized to each architecture's specific resolution and normalized using the ImageNet channel-wise mean and standard deviation. Training was conducted for 15 epochs using the Adam optimizer with a learning rate of 0.0001 and default momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [37]. The categorical cross-entropy loss function was used to optimize multi-class classification performance. A batch size of 32 was used for all experiments, and the model state with the highest validation accuracy was saved for later evaluation. No learning rate scheduling or extra regularization techniques were applied. Training and validation loss and accuracy were systematically monitored after each epoch to track convergence and detect potential overfitting [38].

To comprehensively assess model performance in the multi-class fungal classification task, a set of well-established evaluation metrics was employed. Accuracy quantifies the proportion of correctly classified samples relative to the total number of samples, providing an overall measure of predictive performance. Precision represents the proportion of accurate positive predictions among all positive predictions, reflecting the model's ability to avoid false positives. Recall (or sensitivity) measures the proportion of true positives correctly identified among all actual positive cases, indicating the model's ability to capture relevant instances. F1-Score is the harmonic mean of precision and recall, offering a balanced metric that accounts for both false positives and false

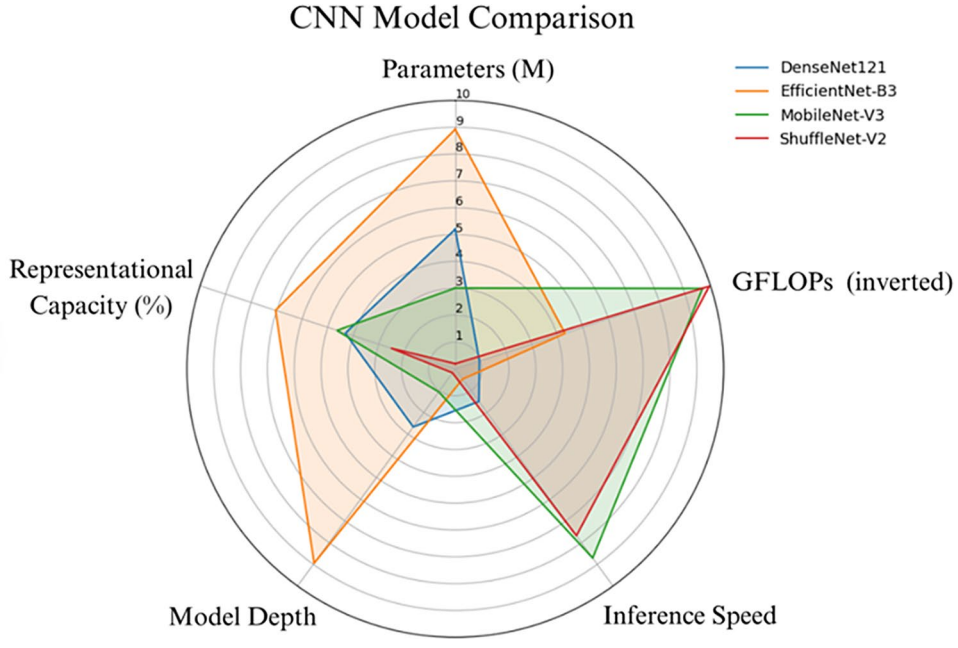

Fig. 2 CNN model comparison across five normalized metrics

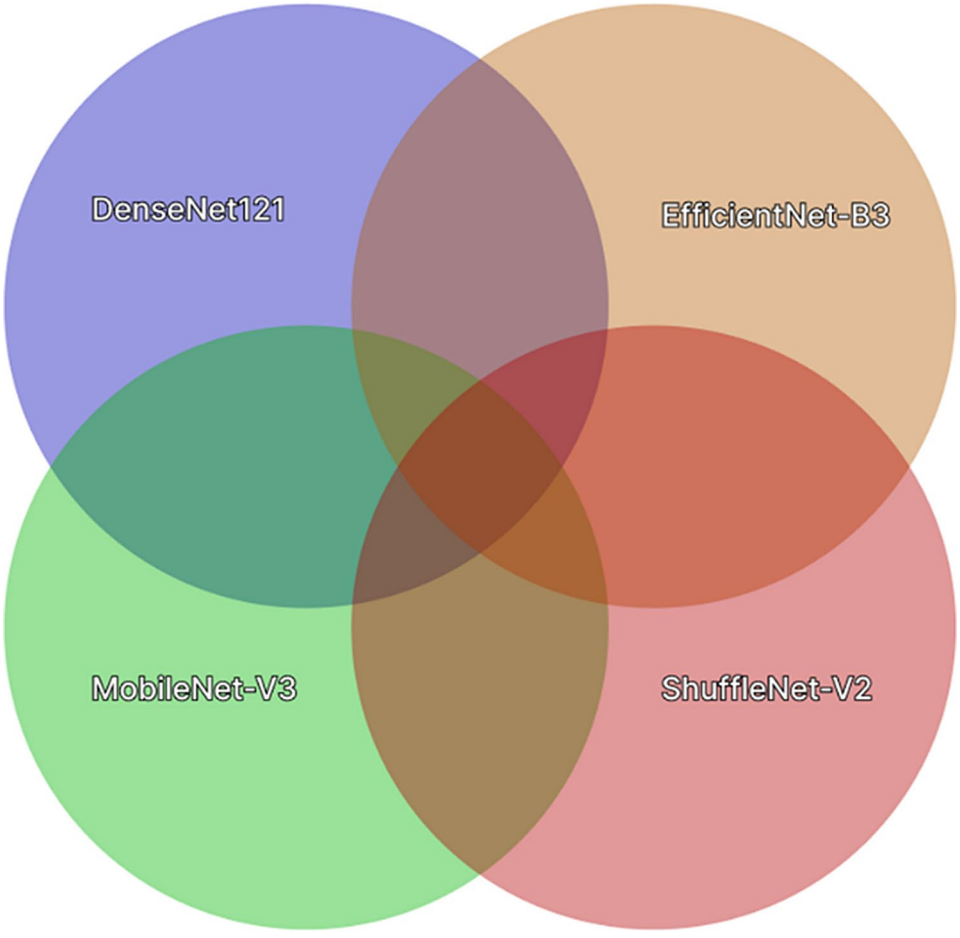

Fig. 3 Venn diagram illustrating the base CNN models and their possible ensemble intersections

negatives. Log Loss evaluates the accuracy of predicted probability distributions rather than hard labels, penalizing confident but incorrect predictions more heavily. Matthews Correlation Coefficient (MCC) is a correlation-based measure that accounts for true and false positives and negatives, making it particularly robust to class imbalance. Finally, the Area Under the Receiver Operating Characteristic Curve (AUC) computed in its macro-averaged form captures the model's ability to discriminate between classes across varying classification thresholds [13–18].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TN + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{4}$$

$$LogLoss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(\hat{y}ij) \tag{5}$$

$$AccessisdeniedAccessisdenied \tag{6}$$

For clarity, the symbols are defined as follows: $AccessisdeniedAccessisdenied$, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively, computed from the confusion matrix; in the multi-class setting, these quantities are obtained in a one-vs-rest manner and aggregated as customary. N represents the total number of evaluated samples. M denotes the total number of classes. $y_{ij} \in \{0, 1\}$ indicates whether sample i belongs to class j (1 if true, 0 otherwise), and $\hat{y}_{ij} \in \{0, 1\}$ is the predicted probability assigned by the model to class j for sample i, with $\sum_{j=1}^{M} \hat{y}_{ij} = 1$ $Precision$ and $Recall$ are

the standard quantities derived from TP, FP, and FN, and the $F1-Score$ is their harmonic mean. Finally, the $MatthewsCorrelationCoefficient\,(MCC)$ is computed from TP, TN, FP, and FN to provide a balanced measure even in the case of class imbalance [13–18].

In addition to the core performance metrics, several auxiliary metrics were calculated for the base models to yield deeper insights into computational efficiency and prediction confidence. $Specificity$ quantifies the proportion of true negatives correctly identified, reflecting the model's capacity to avoid false positives. $Averageinferencetimeperimage$ measures the mean time required by the model to produce a prediction for a single input sample, serving as a proxy for real-time deployment feasibility. $Averagemaximumpredictedprobability$ denotes the mean of the highest-class probabilities across all evaluated samples, providing an aggregate measure of the model's confidence in its predictions. $Averagepredictiveentropy$ captures the uncertainty inherent in the probability distributions output by the model, with lower entropy values indicating more decisive and confident predictions. $Memoryusagepercentage$ reports the average fraction of available system memory consumed during inference, while $GPUmemoryusage$ reflects the peak memory allocation on the GPU in megabytes. Finally, $energyefficiency$ is defined as the number of inferences completed per unit of energy consumed, thereby offering a hardware-level assessment of computational cost-effectiveness.

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$t_{avg} = \frac{1}{N}\sum_{i=1}^{N} t_i \tag{8}$$

$$p_{max,avg} = \frac{1}{N}\sum_{=1}^{N} \max_{1\le j\le M} \hat{y}_{ij} \tag{9}$$

$$AccessisdeniedAccessisdenied \tag{10}$$

$$AccessisdeniedAccessisdenied \tag{11}$$

$$\eta_{energy} = \frac{NumberofInferences}{EnergyConsumption\,(Joules)} \tag{12}$$

The notation introduced for the auxiliary metrics is as follows: t_i denotes the inference time in seconds for the $i-th$ input sample. $\max_{1\le i\le M} \hat{y}_{ij}$ represents the maximum predicted probability among all classes for sample i, and H_{avg} corresponds to the average Shannon entropy of the predicted probability distributions. η_{energy} quantifies the number of inferences performed per joule of energy consumed, providing a hardware-level measure of computational cost-effectiveness [13–18].

These auxiliary metrics were computed exclusively for base models, as their values are inherently tied to the characteristics of individual architectures and their hardware-level execution. In the case of ensemble configurations, such quantities cannot be meaningfully measured, since the predictions are generated by aggregating the outputs of multiple independently trained models rather than executing a single unified network. As a result, hardware-dependent measures such as inference time, memory usage, and energy efficiency would not represent a direct property of the ensemble itself but rather a combined overhead of heterogeneous components, which varies according to the implementation strategy and execution environment. Therefore, to ensure interpretability and comparability, these metrics were omitted for ensemble models.

To enhance interpretability and visualize the decision-making process of the models, both Gradient-weighted Class Activation Mapping (Grad-CAM) [39] and its improved variant Grad-CAM++ were employed [40]. Grad-CAM generates a coarse localization map by computing the weighted sum of the activation maps from the final convolutional layer, where the weights are derived from the gradients of the target class score with respect to the feature maps. Grad-CAM++ extends this method by introducing a more refined weighting scheme that accounts for pixel-level contributions, enabling higher-resolution and more precise localization of class-discriminative regions [40].

For all base CNN architectures (DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2), the techniques were applied to the final convolutional block, enabling a direct mapping between model activations and the corresponding input image regions. For ensemble configurations, heatmaps were first generated individually for each constituent model and then combined using a weighted summation strategy, where the weights corresponded to the optimal values obtained from the ensemble's validation performance. This approach preserved each model's relative contribution to the final decision, producing a unified visual explanation for the ensemble prediction.

The deep learning and ensemble-based analyses presented in this study were designed to be carried out without requiring advanced high-performance computing infrastructure. Model training and evaluation were conducted on a standard desktop workstation, and, on average, a multi-core CPU (e.g., Intel i7 or AMD Ryzen 7 level), at least 16 GB of system memory, and a mid-range GPU (e.g., NVIDIA RTX 3060 or an equivalent with 8–12

GB VRAM) are sufficient to reproduce this work. In addition, Python-based open-source libraries (PyTorch, Albumentations, NumPy, and scikit-learn) were used, and no proprietary or specialized software infrastructure was required. In this respect, the proposed approach offers an accessible and practical solution for both academic and applied research.

Results

This section presents a comprehensive evaluation of the experimental outcomes, encompassing both quantitative performance assessments and qualitative explainability analyses. First, the training and fine-tuning progress of all base CNN architectures is reported, followed by a detailed comparison of their classification performance across multiple evaluation metrics, including Accuracy, Precision, Recall, F1-Score, Log Loss, and Matthews Correlation Coefficient (MCC). Figure 4 presents the evolution of training accuracy over 15 epochs for the four base CNN architectures, while Figure 5 shows the corresponding training loss curves. The same set of metrics is then used to evaluate the proposed ensemble configurations, enabling a direct comparison with individual models. Subsequently, explainability results obtained via Grad-CAM and Grad-CAM++ are discussed, highlighting the models' attention patterns and their biological plausibility, particularly in challenging inter-class scenarios. The section concludes with an overall interpretation of findings, synthesizing insights from both performance and interpretability perspectives to provide a holistic understanding of the models' behavior.

DenseNet121 achieved the fastest convergence, surpassing 95% accuracy by the third epoch and reaching a peak of 99.57% by epoch 12, with loss decreasing sharply from 3.00 to 0.1242. EfficientNet-B3 exhibited steady improvement, exceeding 80% accuracy by epoch 3 and reaching 93.72% in the final epoch, accompanied by a loss reduction from 3.00 to 0.1128. MobileNet-V3 followed a similar trajectory, attaining 82.30% accuracy by epoch 3 and final accuracy of 93.43%, with loss decreasing from 3.00 to 0.2183. ShuffleNet-V2, while showing consistent gains, lagged behind the others, improving from 17.28% at epoch 1 to 84.04% at epoch 15, and reducing loss from 3.00 to 0.5153. Overall, these results highlight that deeper and more parameter-rich models such as DenseNet121 and EfficientNet-B3 converge faster and achieve higher training performance, while lightweight architectures like ShuffleNet-V2 trade off accuracy for computational efficiency. Figures 6 and 7 provide a comparative overview of the validation accuracy and loss curves obtained for the evaluated models.

Validation accuracy curves in Figure 6 show that all models achieve rapid performance gains in the initial epochs, with DenseNet121 and EfficientNet-B3 reaching the highest accuracy of 93.22% by epoch 15, MobileNet-V3 stabilizing around 91.44%, and ShuffleNet-V2 despite a markedly lower starting point improving steadily to 88.22%.

The corresponding validation loss curves in Figure 7 reveal that EfficientNet-B3 attains the lowest final loss (0.2243), followed closely by DenseNet121 (0.2782) and MobileNet-V3 (0.2840), while ShuffleNet-V2 exhibits a slower yet consistent decline to 0.3696. Overall, these

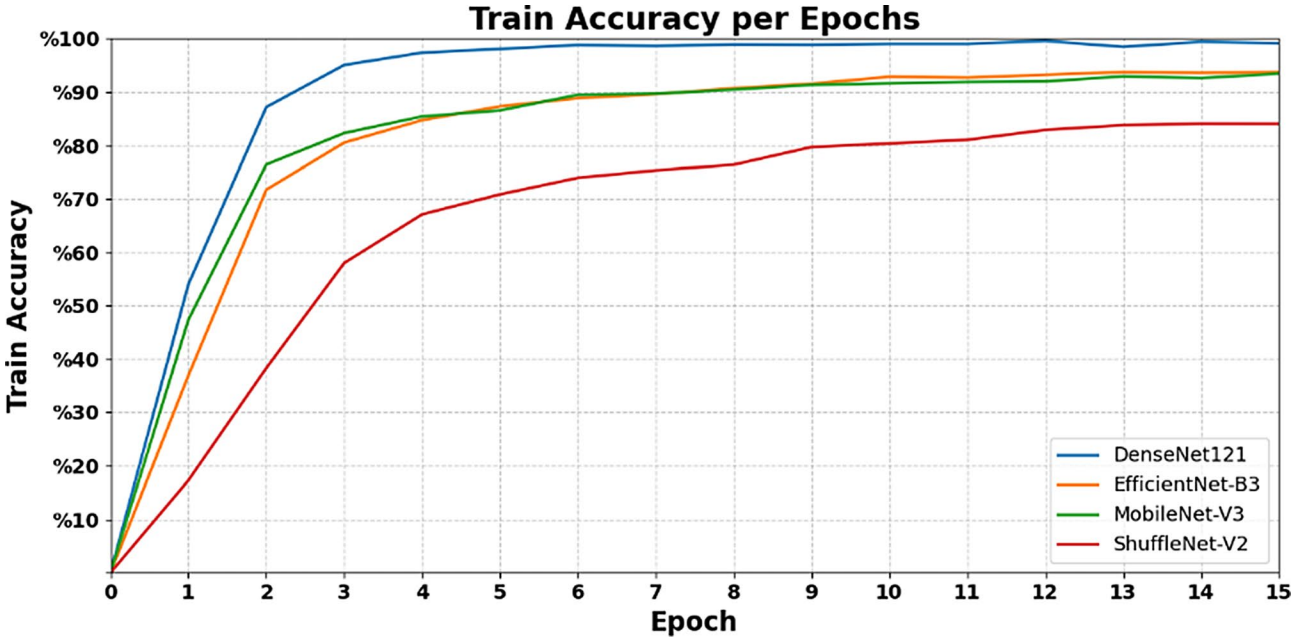

Fig. 4 Training accuracy curves of four base CNN models over 15 epochs

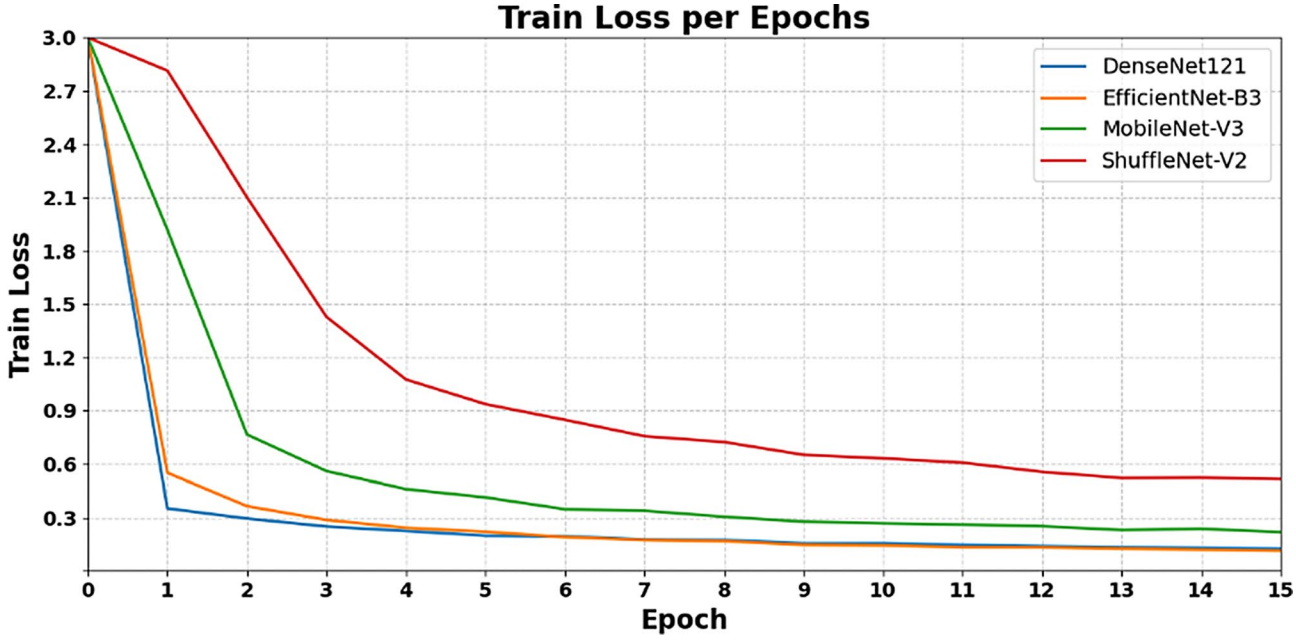

Fig. 5 Training loss curves of four base CNN models over 15 epochs

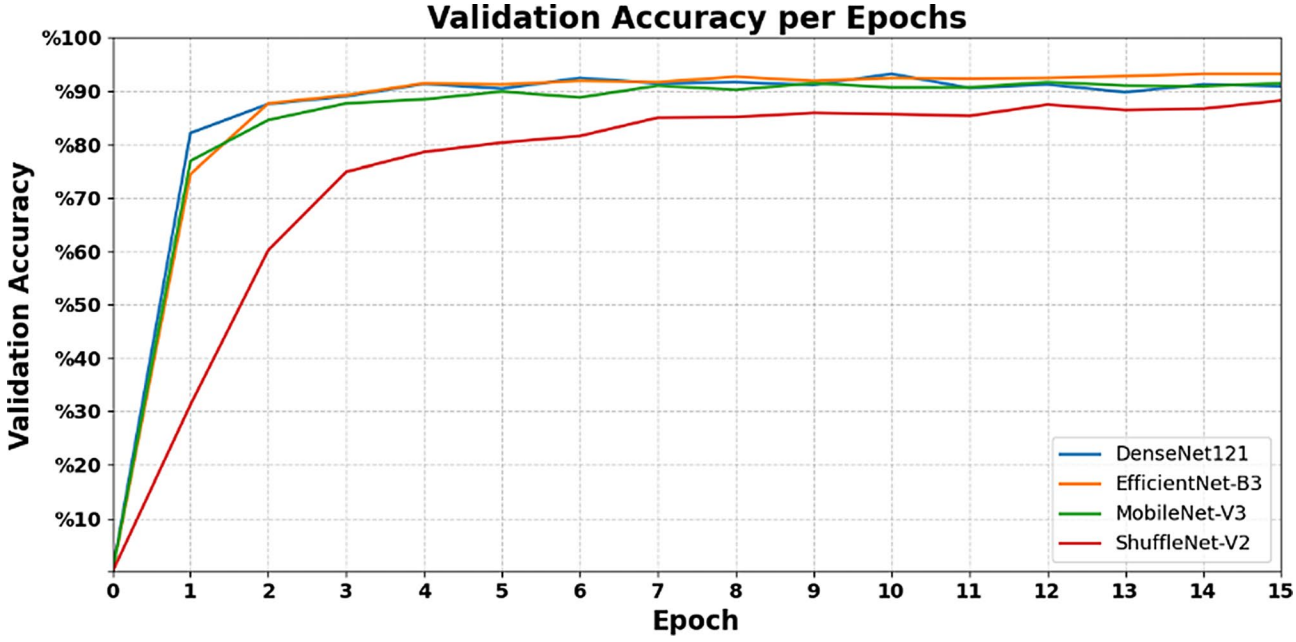

Fig. 6 Validation accuracy curves of four base CNN models over 15 epochs

results indicate that DenseNet121 and EfficientNet-B3 achieve the most favorable trade-off between validation accuracy and loss. In Fig. 8, the test set accuracy scores of the individual models are presented, while Figure 9 displays the accuracy results of the ensemble models constructed with different weight configurations.

Among the single models, EfficientNet-B3 achieved the highest accuracy at 93.0%, followed by MobileNet-V3 at 91.0%, DenseNet121 at 90.0%, and ShuffleNet-V2 at 86.78%.

In the ensemble configurations, all tested combinations surpassed the accuracy of their constituent single models, confirming the performance gains from model integration. Notably, several ensembles involving EfficientNet-B3 consistently delivered high accuracy: the DenseNet121 + EfficientNet-B3 pair reached 95.0%, while EfficientNet-B3 + MobileNet-V3 attained 94.56%, and

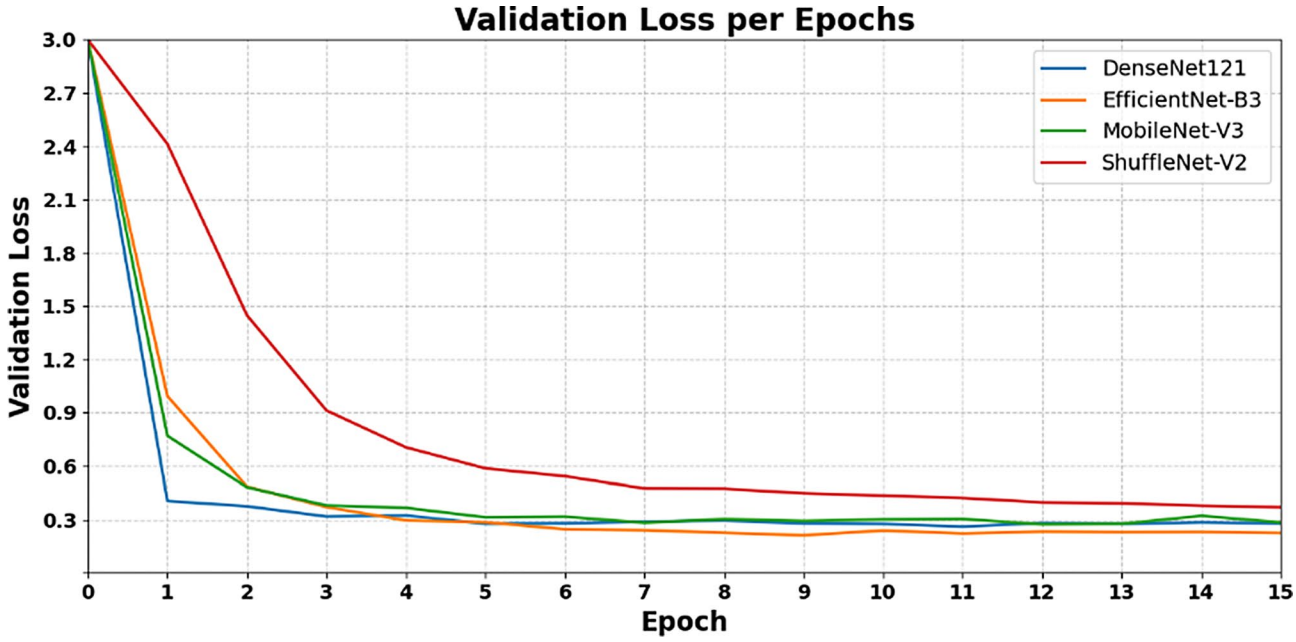

Fig. 7 Validation loss curves of four base CNN models over 15 epochs

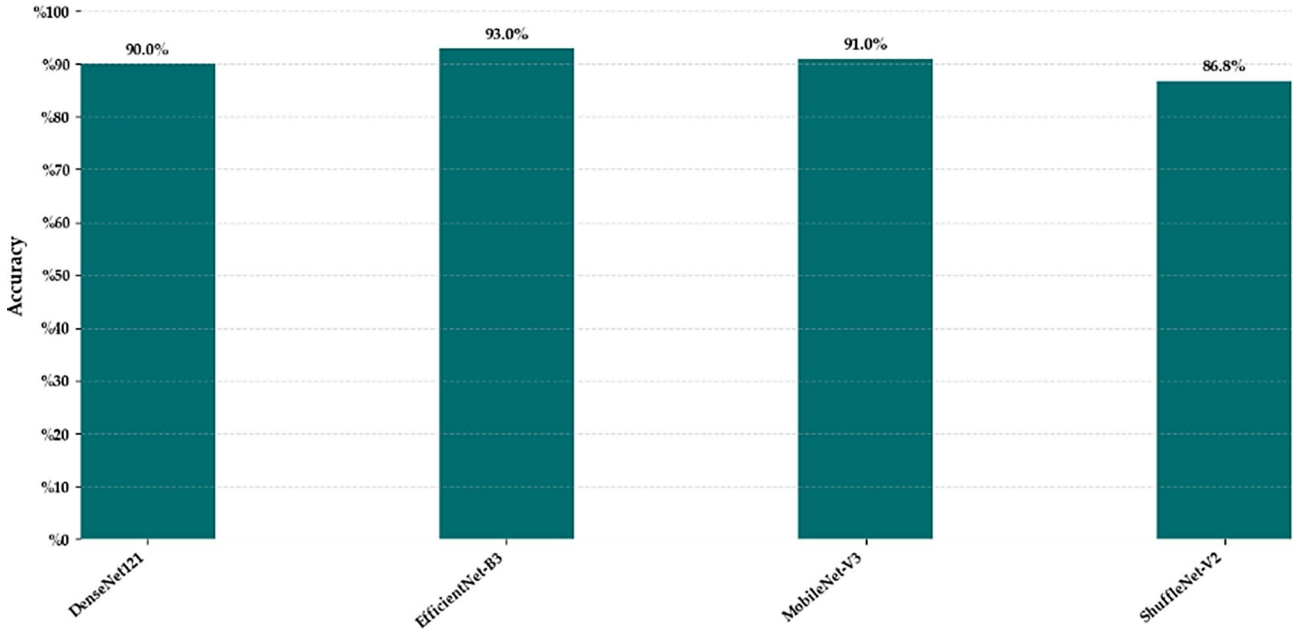

Fig. 8 Accuracy scores for base CNN architectures

EfficientNet-B3 + ShuffleNet-V2 achieved 94.44%. Three-model ensembles also showed strong results, such as DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 (95.44%) and DenseNet121 + EfficientNet-B3 + MobileNet-V3 (95.33%). The highest performance was recorded by the four-model ensemble (DenseNet121, EfficientNet-B3, MobileNet-V3, ShuffleNet-V2) at (95.67%). These results indicate that while the largest ensemble achieved the top accuracy, several smaller ensembles approached its performance, suggesting that well-selected subsets of complementary architectures can yield nearly optimal results with potentially reduced computational cost. The precision results depicted in Figure 10 for single models and Figure 11 for ensemble configurations reveal a consistent performance enhancement when leveraging model combinations.

Among the individual models, EfficientNet-B3 achieved the highest precision (93.20%), followed closely by MobileNet-V3 (91.50%) and DenseNet121 (90.51%), while ShuffleNet-V2 trailed with (87.17%). These results

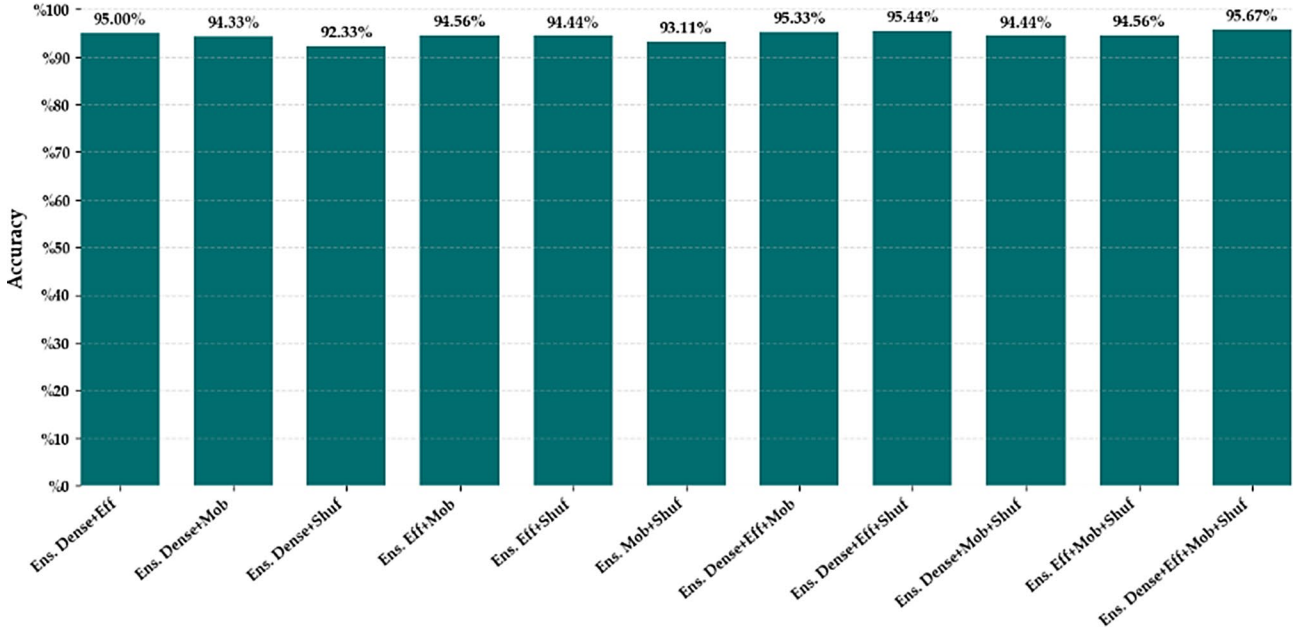

Fig. 9 Accuracy scores for ensemble model configurations

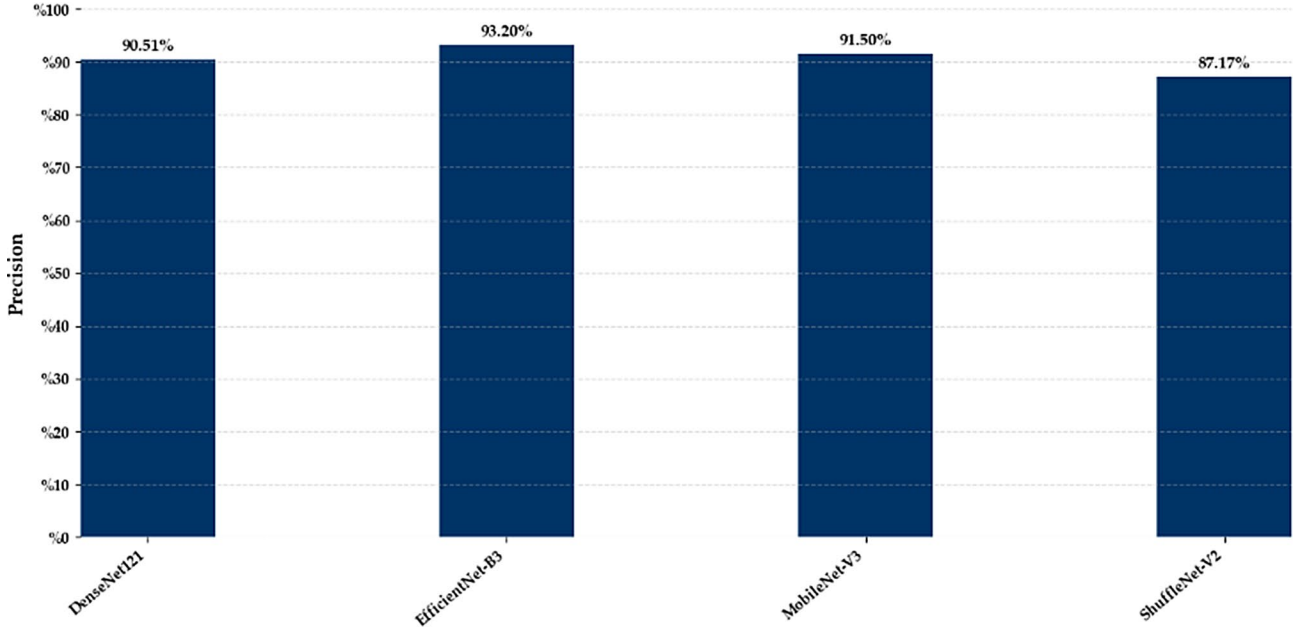

Fig. 10 Precision scores for base CNN architectures

indicate that, in isolation, EfficientNet-B3 demonstrates superior ability to minimize false positives across classes.

Ensemble methods exhibit notable improvements across nearly all configurations, with precision values exceeding 92% in all cases. The highest precision was recorded for the four-model ensemble combining DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2 (95.78%), closely followed by DenseNet121 + EfficientNet-B3 + MobileNet-V3 (95.53%) and DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 (95.63%). Even the weakest ensemble configuration, DenseNet121 + ShuffleNet-V2, outperformed the least precise single model by over 5% points. This consistent superiority across ensemble setups highlights the robustness of multi-model strategies in reducing false positive rates and ensuring more reliable classification decisions, particularly in challenging cases where individual models may err. Recall results in Figure 12 (single models) and Figure 13 (ensembles) show consistent gains from model aggregation.

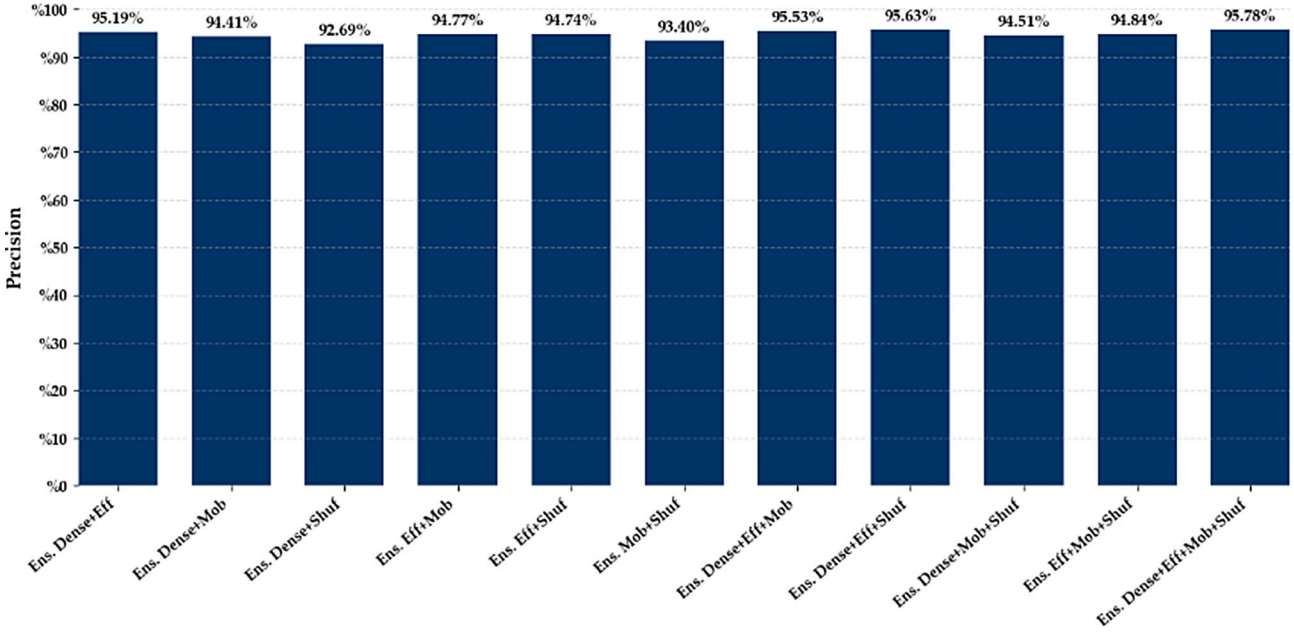

Fig. 11 Precision scores for ensemble model configurations

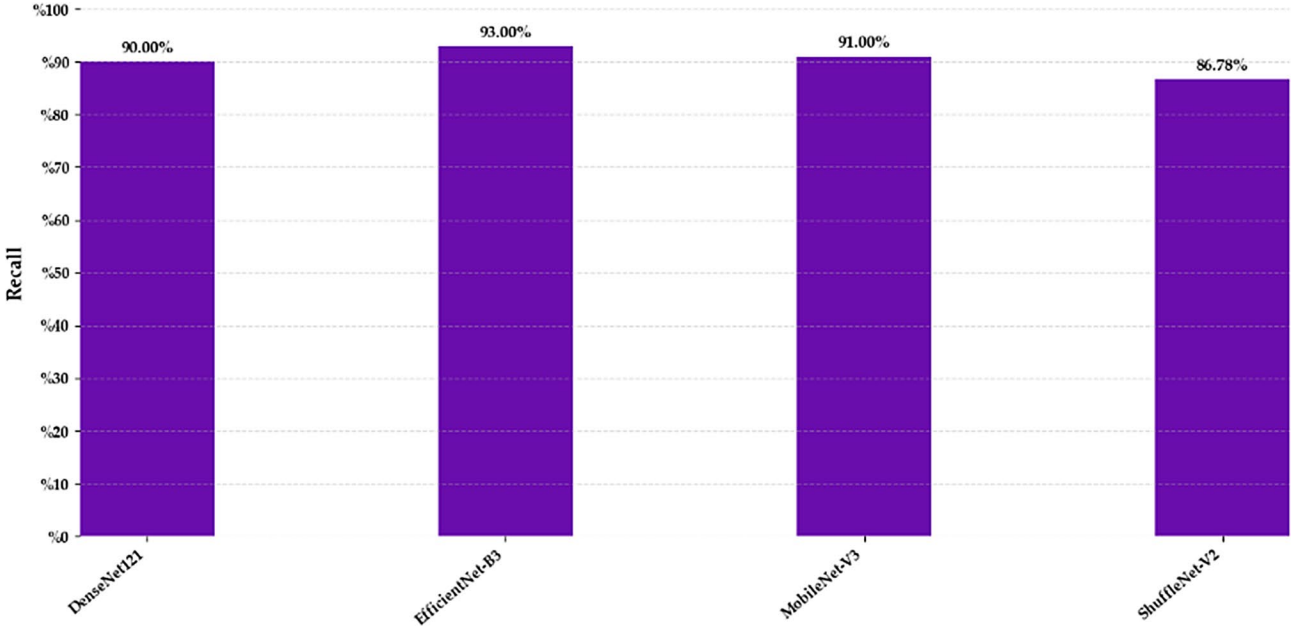

Fig. 12 Recall scores for base CNN architectures

Among individual networks, EfficientNetB3 attains the highest recall at 93.0%, followed by MobileNetV3 (91.0%) and DenseNet121 (90.0%), with ShuffleNetV2 trailing at 86.78%.

All ensemble configurations improve sensitivity, with recalls ≥ 92.33% across the board. Strong two and threemodel combinations include DenseNet121 + EfficientNetB3 (95.00%), EfficientNetB3 + MobileNetV3 (94.56%), EfficientNetB3 + ShuffleNetV2 (94.44%), DenseNet121 + EfficientNetB3 + MobileNetV3 (95.33%), and DenseNet121 + EfficientNetB3 + ShuffleNetV2 (95.44%). The fourmodel ensemble (DenseNet121 + EfficientNetB3 + MobileNetV3 + ShuffleNetV2) achieves the best recall at 95.67%. Notably, even the weakest ensemble (DenseNet121 + ShuffleNetV2, 92.33%) surpasses the least sensitive single model by ≈5.5% points and exceeds DenseNet121 alone, indicating that complementary error patterns reduce false negatives across species. Figures 14 and 15 present the F1-score results for individual models and ensemble configurations, respectively.

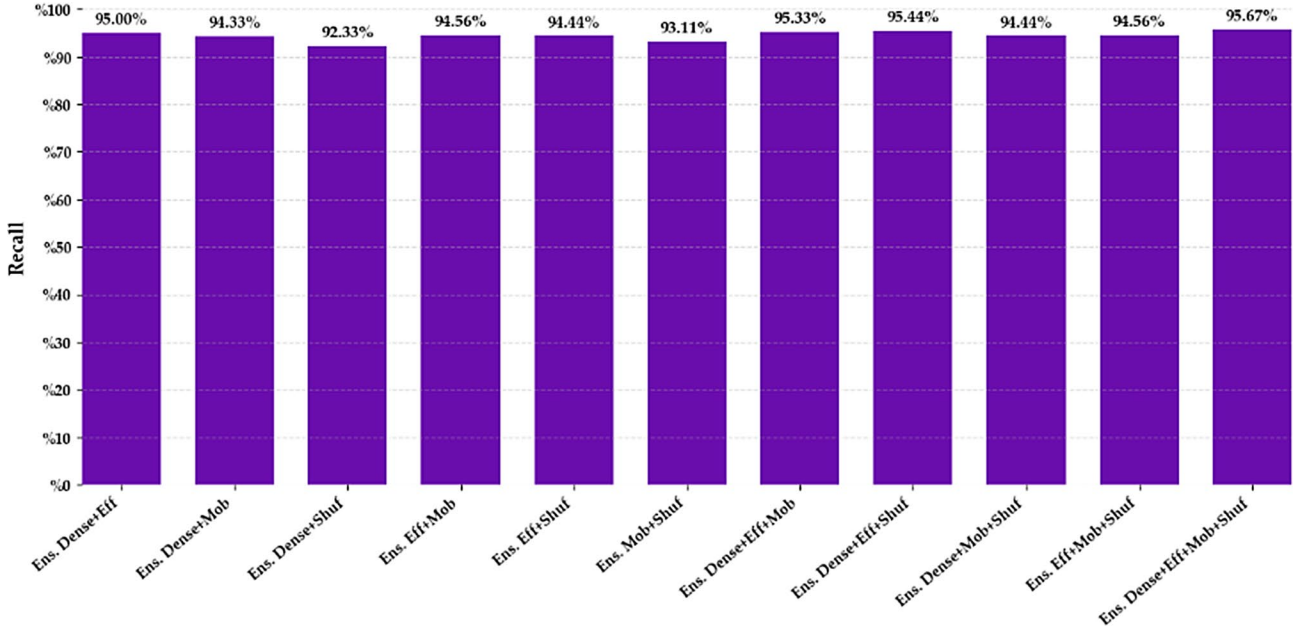

Fig. 13 Recall scores for ensemble model configurations

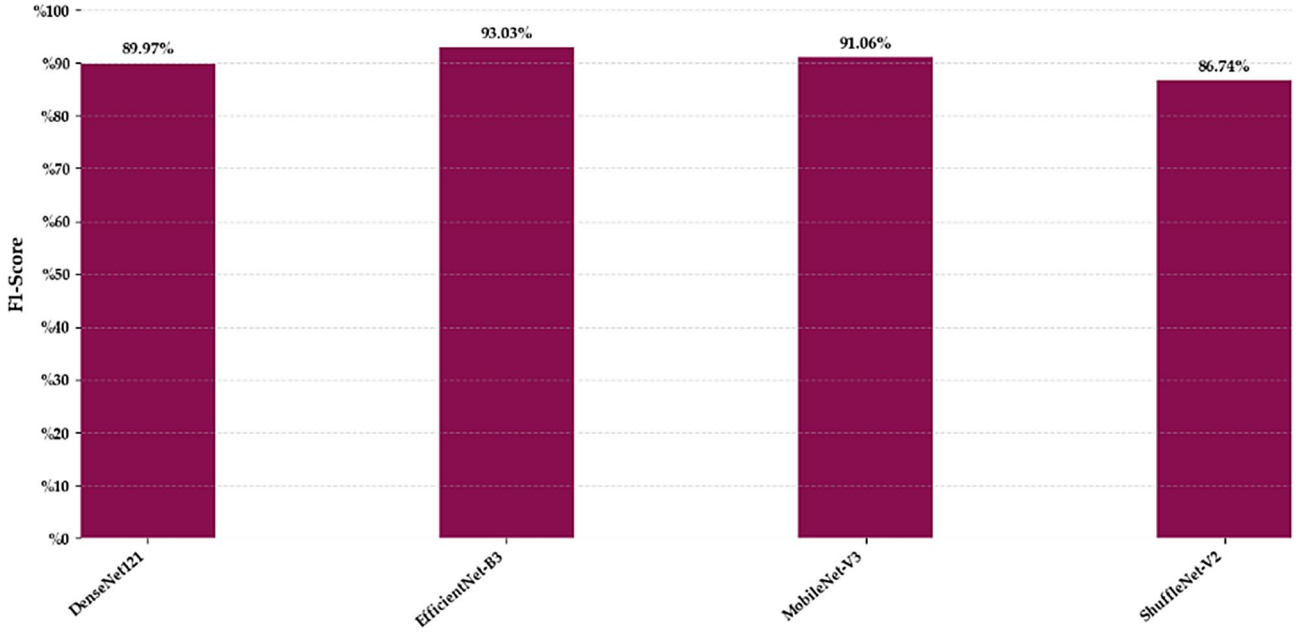

Fig. 14 F1-scores for base CNN architectures

The F1-score results presented in Figure 14 for individual models indicate that EfficientNet-B3 achieved the highest balance between precision and recall with an F1-score of 93.03%, followed by MobileNet-V3 at 91.06% and DenseNet121 at 89.97%. ShuffleNet-V2 recorded the lowest value at 86.74%, reflecting its relatively lower classification performance across classes.

Figure 15, illustrating ensemble configurations, shows that combining multiple architectures consistently improved F1-scores compared to single models. The top-performing ensemble, integrating DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2 with respective weights of 0.30, 0.40, 0.20, and 0.10, achieved an F1-score of 95.67%, marking the highest among all tested configuration-sall_results. Other high-ranking ensembles included DenseNet121 + EfficientNet-B3 + MobileNet-V3 (95.35%) and DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 (95.45%). Notably, even the lowest-performing ensemble, DenseNet121 + ShuffleNet-V2, with an F1-score of

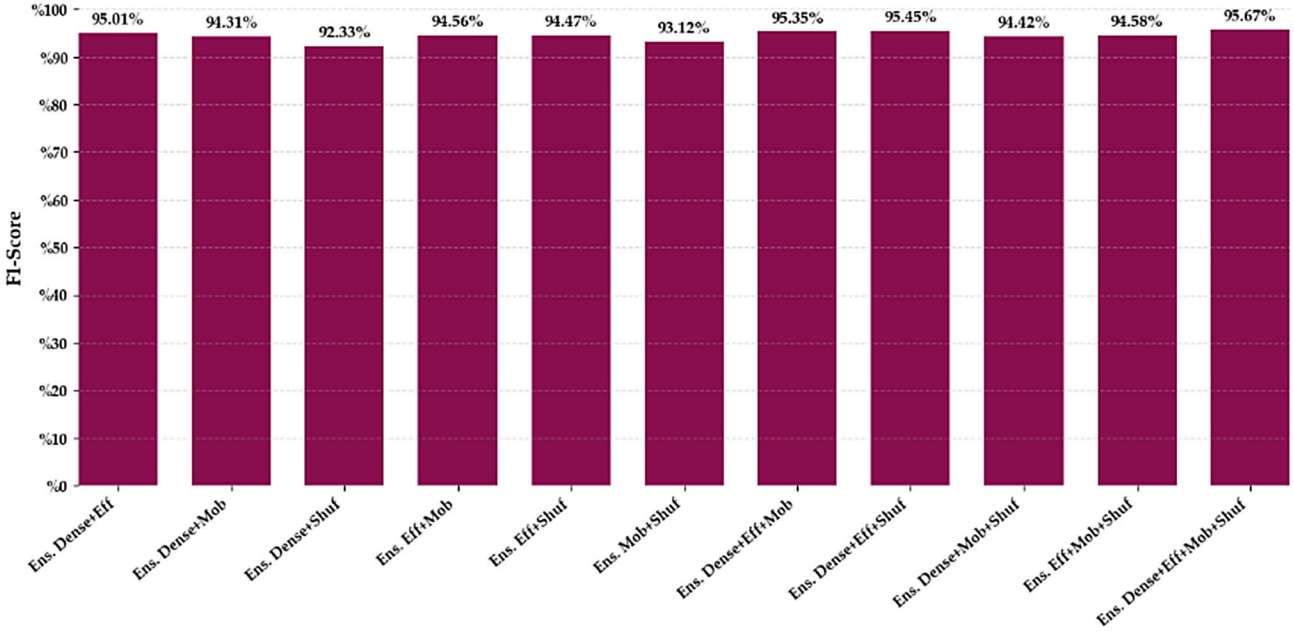

Fig. 15 F1-scores for ensemble model configurations

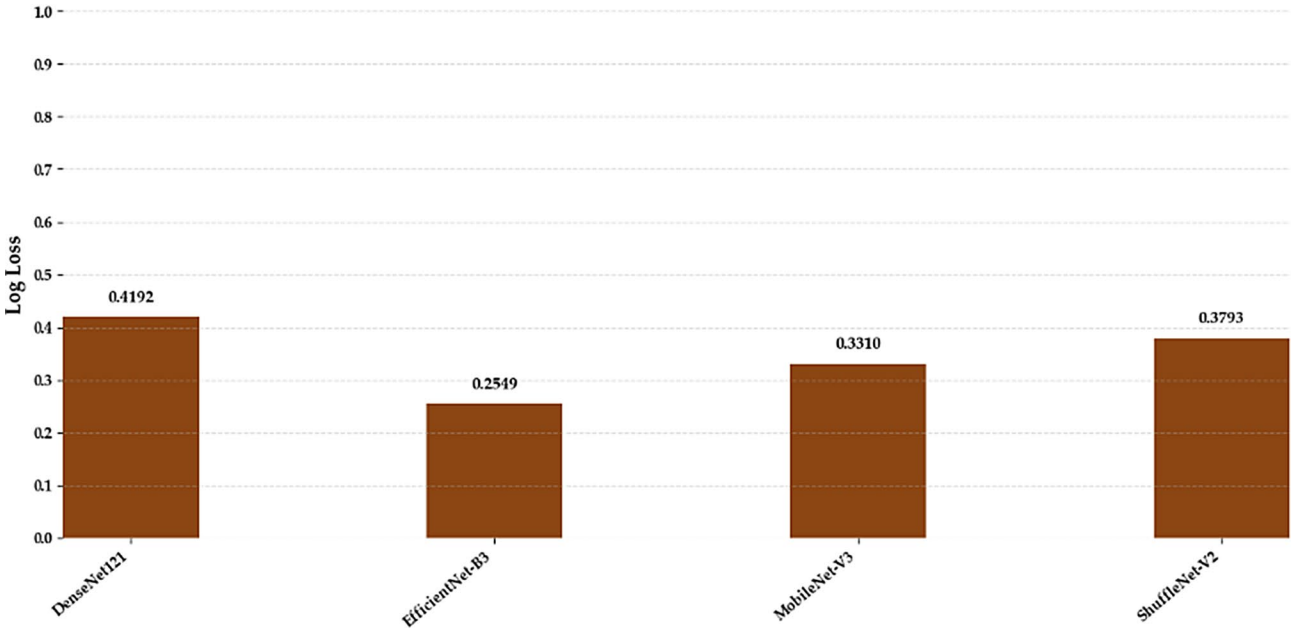

Fig. 16 Log loss scores for base CNN architectures

92.33%, still outperformed the weakest individual model, underscoring the robustness and performance stability gained through ensemble strategies. The log loss results in Figs. 16 and 17 provide additional insight into the calibration and confidence of model predictions.

Among single models, EfficientNet-B3 achieved the lowest log loss (0.2549), indicating well-calibrated confidence estimates, while DenseNet121 recorded the highest (0.4192), suggesting more overconfident or miscalibrated predictions despite its reasonable accuracy. MobileNet-V3 (0.3310) and ShuffleNet-V2 (0.3793) fell in between, reflecting moderate calibration quality.

For ensemble architectures, a clear reduction in log loss was observed across most combinations compared to their individual counterparts. The best performance was achieved by the four-model ensemble with weights 0.30 (DenseNet121), 0.40 (EfficientNet-B3), 0.20 (MobileNet-V3), and 0.10 (ShuffleNet-V2), reaching the lowest log loss of 0.1751. Other high-performing configurations included DenseNet121 + EfficientNet-B3 + MobileNet-V3

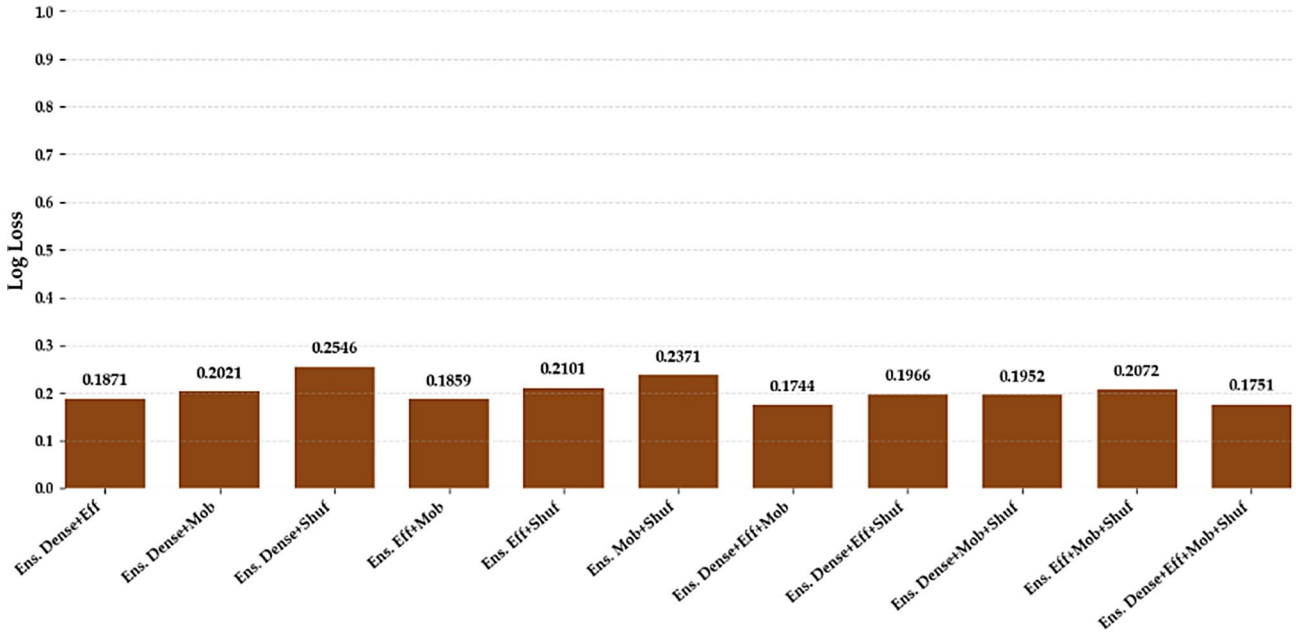

Fig. 17 Log loss scores for ensemble model configurations

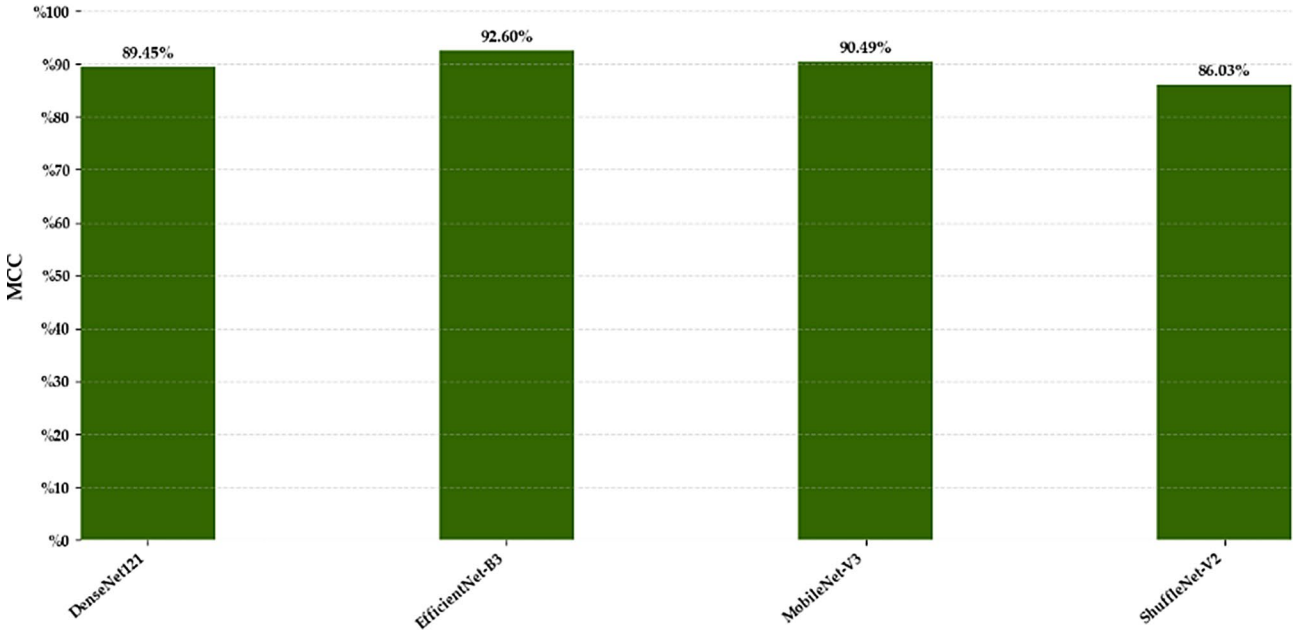

Fig. 18 MCC scores for base CNN architectures

(0.1744) and EfficientNet-B3 + MobileNet-V3 (0.1859), all substantially outperforming the best single model. These results confirm that ensembling not only boosts classification accuracy and F1-score but also leads to better-calibrated probability estimates, enhancing model reliability in practical decision-making scenarios. Figure 18 illustrates the MCC performance of the single models, while Figure 19 presents the MCC values for the ensemble configurations.

Across all models, ensemble methods consistently outperformed their single-model counterparts. Among the single models in Fig. 18, EfficientNet-B3 achieved the highest MCC at 92.60%, followed by MobileNet-V3 (90.49%) and DenseNet121 (89.45%), with ShuffleNet-V2 trailing at 86.03%.

In Fig. 19, all ensemble configurations surpassed 91.90%, with the DenseNet121 + EfficientNet-B3 + MobileNet-V3 + ShuffleNet-V2 combination achieving the top score of 95.42%, closely followed by

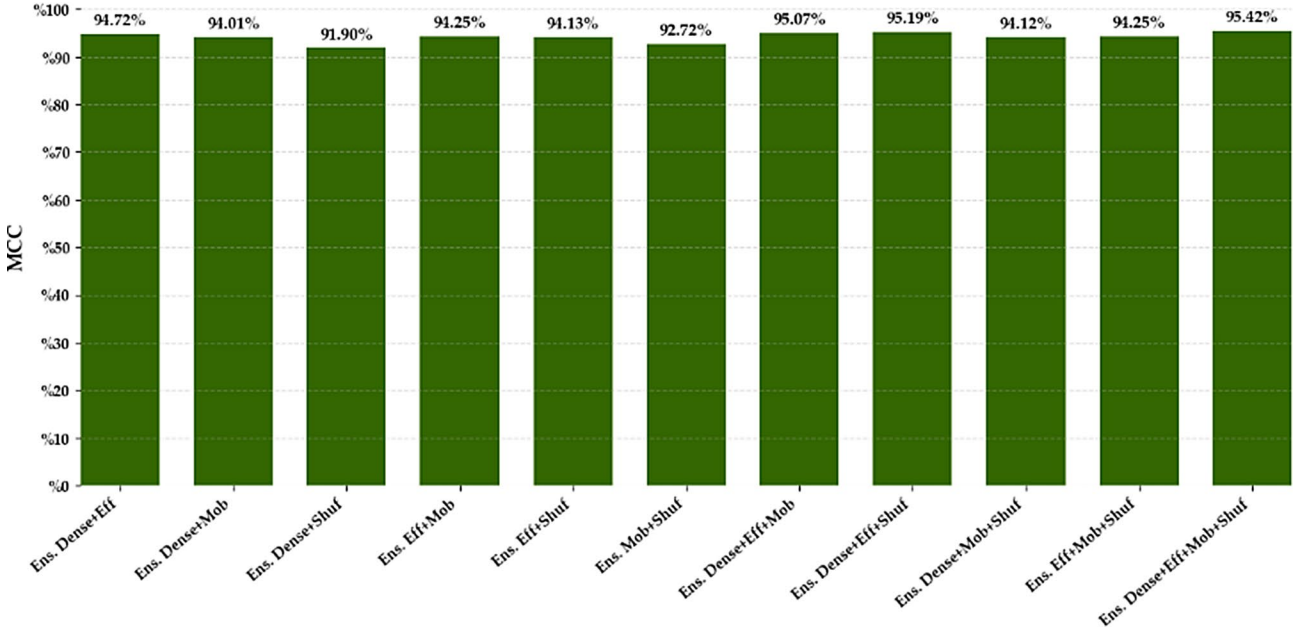

Fig. 19 MCC scores for ensemble model configurations

DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 (95.19%) and DenseNet121 + EfficientNet-B3 + MobileNet-V3 (95.07%). These results highlight the consistent advantage of ensemble strategies in maintaining balanced predictive performance across all classes, a particularly valuable attribute in multiclass mushroom classification where interclass similarity is high. Specificity measures the proportion of actual negative instances correctly identified by a model. For single models, this metric was directly computed since each prediction was assigned to a discrete class label. In contrast, in ensemble architectures, final predictions were obtained through a weighted averaging of the output probabilities from the individual models. Under this approach, class-specific true negative (TN) and false positive (FP) counts could not be directly determined, as predictions were aggregated at the probability level rather than assigned via binary (0/1) classification. In multi-class settings, this occasionally resulted in cases $TN + FP = 0$, yielding undefined $(-\infty)$ specificity values.To preserve methodological consistency, the specificity analysis was therefore performed exclusively on single models, as shown in Fig. 20.

Figure 20 compares the specificity performance of these models, with EfficientNet-B3 achieving the highest score (93.20%), followed by MobileNet-V3 (91.50%), DenseNet121 (90.51%), and ShuffleNet-V2 (87.17%). These findings suggest that deeper and more complex architectures tend to achieve superior performance in distinguishing negative classes. The comparative results in the Table 2 reveal that ensemble models consistently surpass individual models across nearly all metrics.

Among the single models, EfficientNet-B3 achieved the highest performance (93.00% accuracy, 99.70% AUC, MCC 92.60), followed by MobileNet-V3 and DenseNet121, while ShuffleNet-V2 recorded the lowest scores. Pairwise ensembles demonstrated notable improvements over their constituent models, with DenseNet121 + EfficientNet-B3 (95.00% accuracy, MCC 94.72) and EfficientNet-B3 + MobileNet-V3 (94.56% accuracy, MCC 94.25) showing strong results. Triple ensembles further enhanced performance, particularly DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 (95.44% accuracy) and DenseNet121 + EfficientNet-B3 + MobileNet-V3 (95.33% accuracy), both delivering higher MCC and lower log loss values. The four-model ensemble achieved the best overall performance, with 95.67% accuracy, the highest MCC (95.42), and the lowest log loss (17.51), indicating superior predictive reliability and robustness. These findings highlight the advantage of combining complementary models, with performance gains becoming more pronounced as ensemble size increases. As shown in Table 3, the individual models are compared in terms of specificity, inference time, average maximum probability, and average entropy.

The results show that EfficientNet-B3 achieved the highest specificity (93.20%) and the highest average maximum probability (94.70%) while also recording the lowest average entropy (17.44), indicating strong confidence and reliability in its predictions. MobileNet-V3 and DenseNet121 demonstrated comparable specificity values (91.50% and 90.51%, respectively) and moderate average entropy, but MobileNet-V3 was significantly faster in inference time (0.77s vs. 2.34s). ShuffleNet-V2

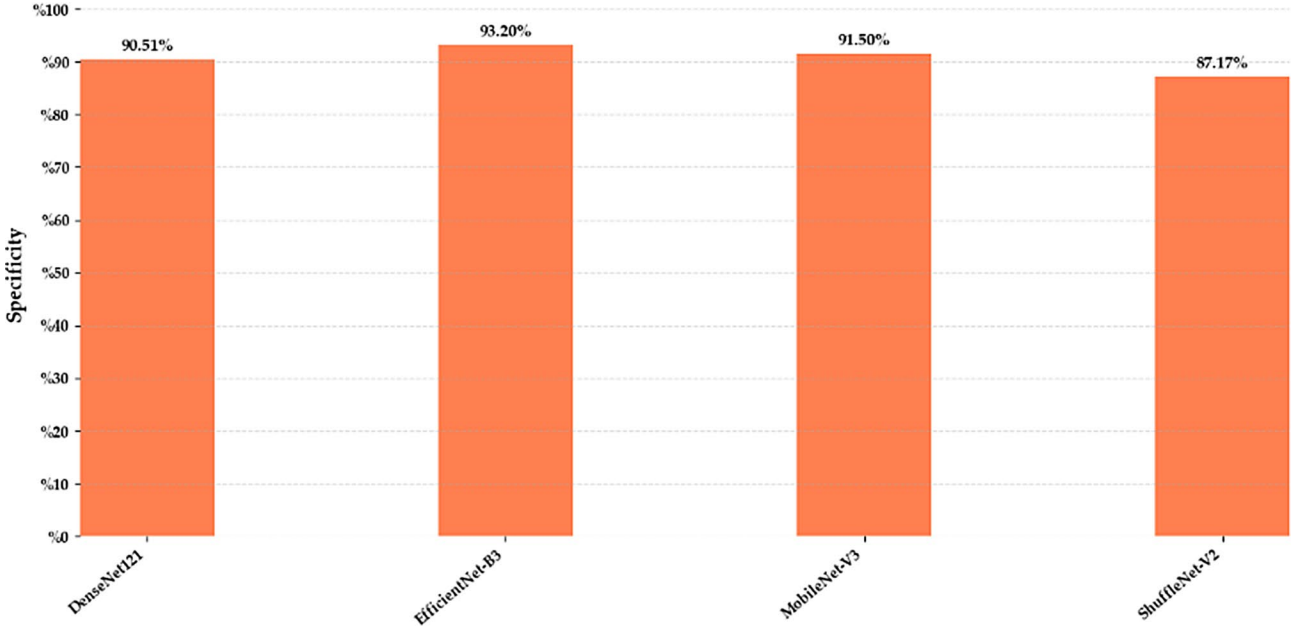

Fig. 20 Specificity scores for base CNN architectures

Table 2 Performance metrics of individual and ensemble models on the test dataset

Models	Accuracy	Precision	Recall	F1-Score	AUC	MCC	Log Loss
DenseNet121	90.00	90.51	90.00	89.97	99.24	89.45	41.92
EfficientNet-B3	**93.00**	**93.20**	**93.00**	**93.03**	**99.70**	**92.60**	**25.49**
MobileNet-V3	91.00	91.50	91.00	91.06	99.65	90.49	33.10
ShuffleNet-V2	86.78	87.17	86.78	86.74	99.53	86.03	37.93
Dense+Eff	**95.00**	**95.19**	**95.00**	**95.01**	**99.86**	**94.72**	**18.71**
Dense+Mob	94.33	94.41	94.33	94.31	99.85	94.01	20.21
Dense+Shuf	92.33	92.69	92.33	92.33	99.78	91.90	25.46
Eff+Mob	94.56	94.77	94.56	94.56	99.88	94.25	18.59
Eff+Shuf	94.44	94.74	94.44	94.47	99.82	94.13	21.01
Mob+Shuf	93.11	93.40	93.11	93.12	99.84	92.72	23.71
Dense+Eff+Mob	95.33	95.53	95.33	95.35	99.89	95.07	17.44
Dense+Eff+Shuf	**95.44**	**95.63**	**95.44**	**95.45**	**99.85**	**95.19**	**19.66**
Dense+ Mob+Shuf	94.44	94.51	94.44	94.42	99.88	94.12	19.52
Eff+Mob+Shuf	94.56	94.84	94.56	94.58	99.86	94.25	20.72
Dense+Eff+Mob+Shuf	**95.67**	**95.78**	**95.67**	**95.67**	**99.90**	**95.42**	**17.51**

Table 3 Specificity, inference time, average maximum probability, and average entropy for individual models

Models	Specificity	Inference Time	Average Max Prob	Average Entropy
DenseNet121	90.51	2.34	93.98	19.49
EfficientNet-B3	**93.20**	**1.95**	**94.70**	**17.44**
MobileNet-V3	91.50	0.77	93.35	19.30
ShuffleNet-V2	87.17	0.79	87.43	38.52

had the lowest specificity (87.17%) and the lowest average maximum probability (87.43%), along with a substantially higher average entropy (38.52), reflecting lower predictive certainty despite having a relatively short inference time (0.79s). Overall, EfficientNet-B3 stands out as the most balanced model, offering high specificity and prediction confidence with reasonable inference speed, whereas ShuffleNet-V2 exhibits a trade-off between speed and predictive reliability. Figures 23–24 illustrate Grad-CAM and Grad-CAM++ maps, revealing model-specific attention patterns for the selected mushroom species.

The Grad-CAM and Grad-CAM++ visualizations for the four individual models DenseNet121, EfficientNet-B3, MobileNet-V3, and ShuffleNet-V2 demonstrate distinct attention patterns when classifying *Amanita muscaria*, *Galerina marginata*, and *Mycena pura*. DenseNet121 exhibits a concentrated focus on biologically relevant

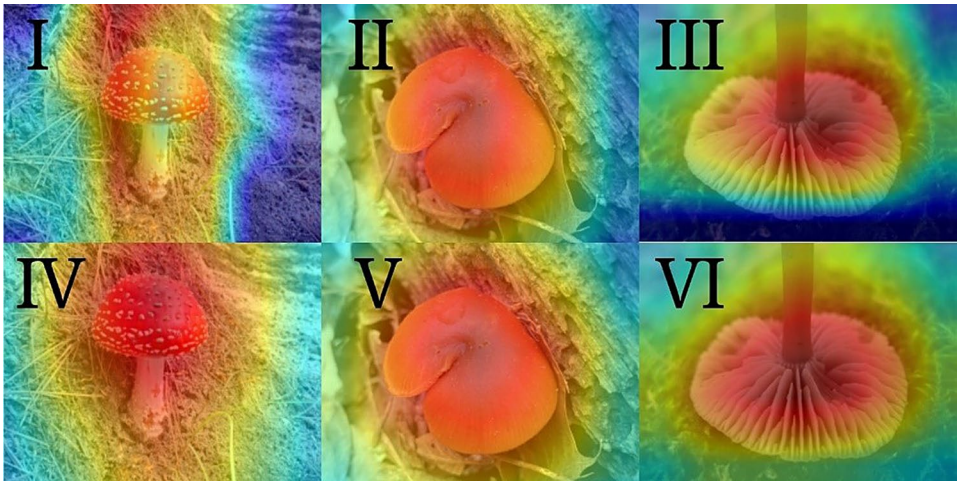

Fig. 21 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the individual model EfficientNet-B3, highlighting the discriminative regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

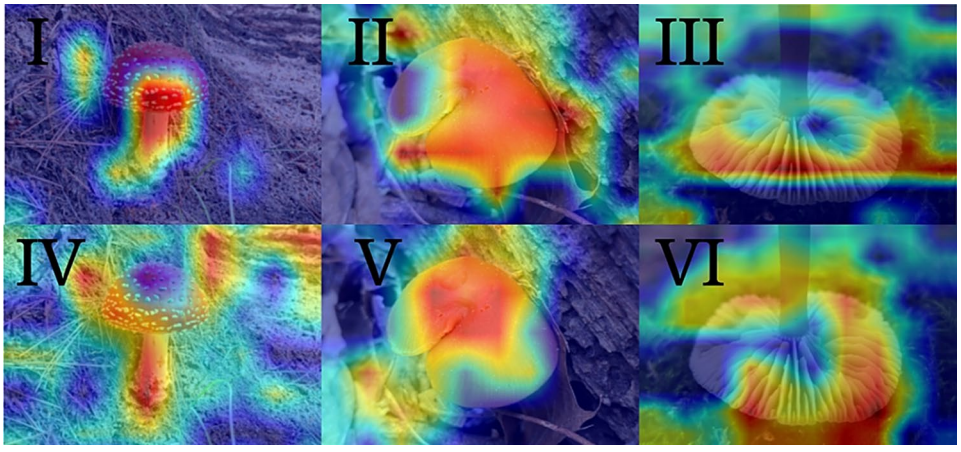

Fig. 22 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the individual model MobileNet-V3, highlighting the discriminative regions for *Amanita muscaria* (i, IV), *Galerina marginata* (ii, V), and *Mycena pura* (iii, VI)

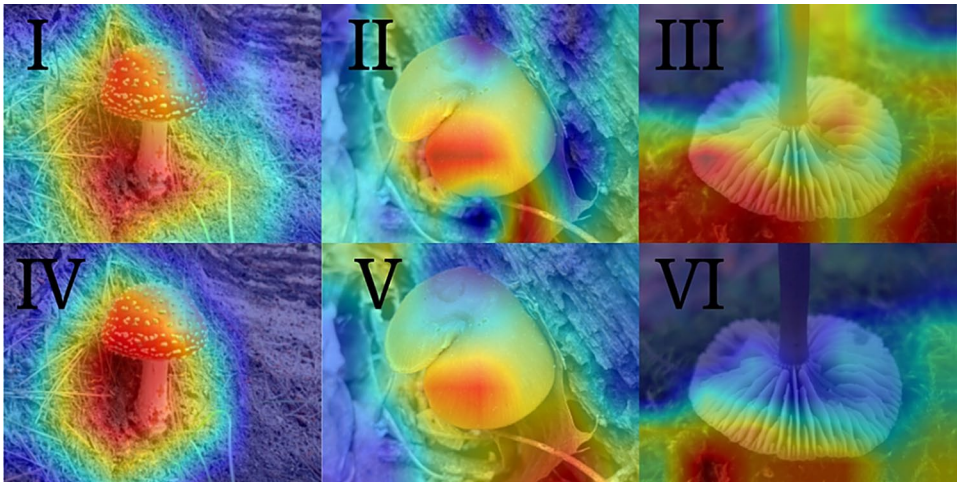

Fig. 23 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the individual model DenseNet121, highlighting the discriminative regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

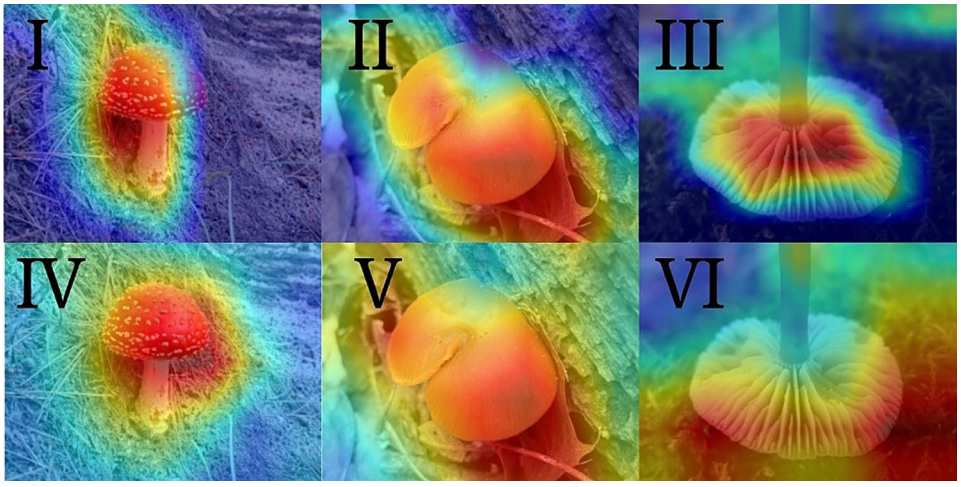

Fig. 24 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the individual model ShuffleNet-V2, highlighting the discriminative regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

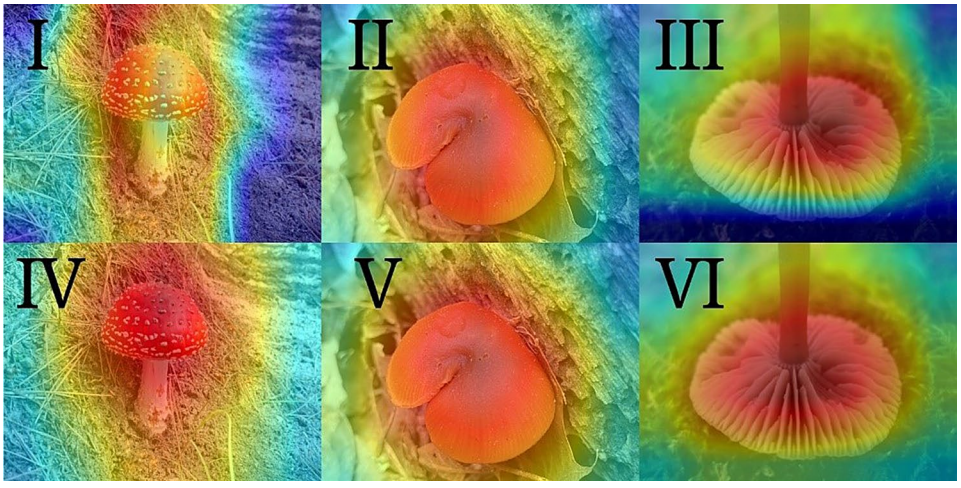

Fig. 25 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model Densenet121 + MobileNet-V3, showing the attention focus for *Amanita muscaria* (i, IV), *Galerina marginata* (ii, V), and *Mycena pura* (iii, VI)

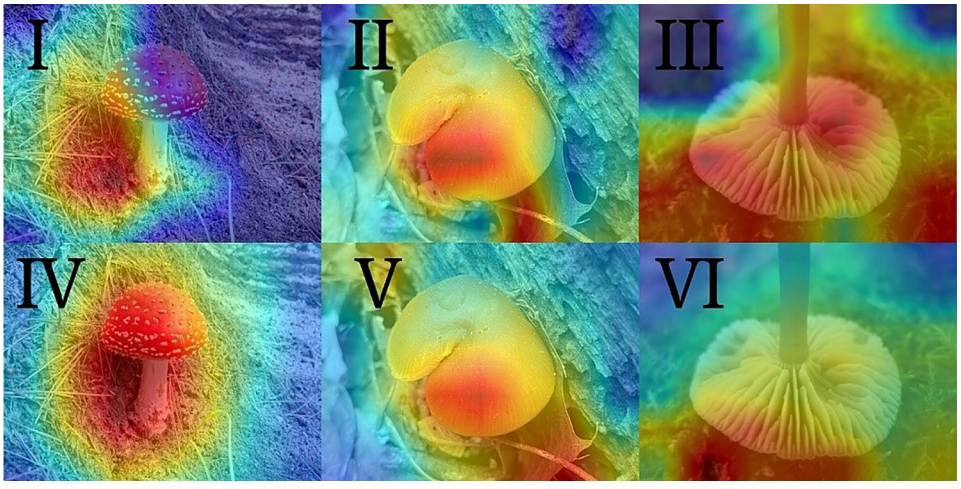

Fig. 26 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model Densenet121 + ShuffleNet-V2, showing the attention focus for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

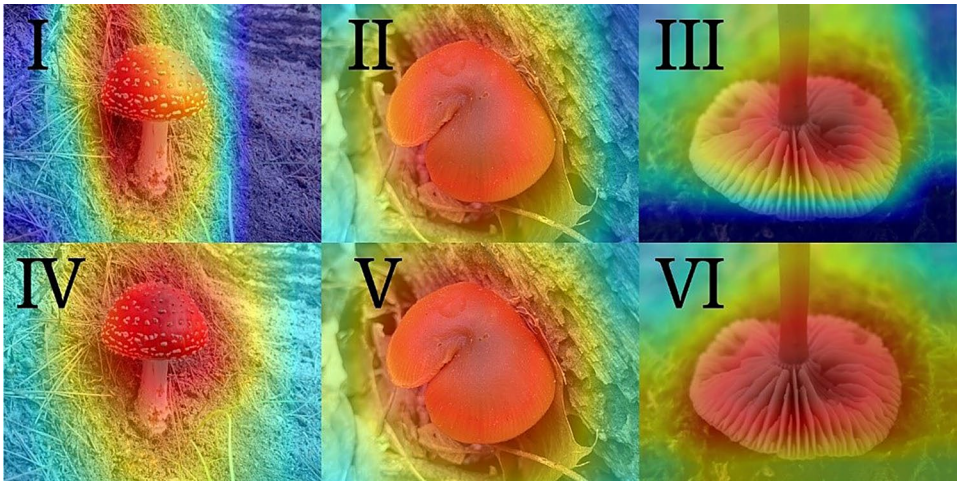

Fig. 27 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model EfficientNet-B3 + MobileNet-V3, showing the attention focus for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

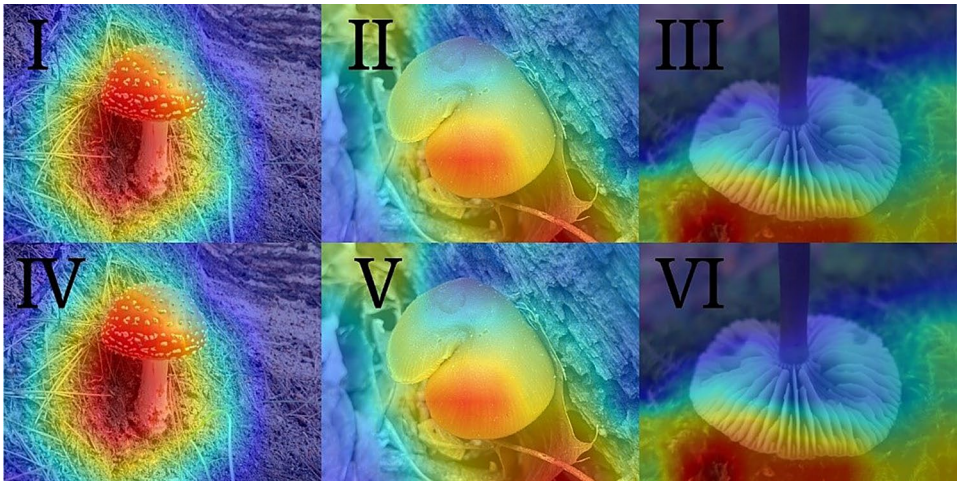

Fig. 28 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model EfficientNet-B3 + ShuffleNet-V2, showing the attention focus for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

regions, such as the pileus and stipe, with Grad-CAM++ providing sharper localization around gill structures. EfficientNet-B3 shows a balanced attention distribution, effectively capturing both global shape and fine-grained textural details, with Grad-CAM++ enhancing boundary precision. MobileNet-V3, while lightweight, maintains targeted attention on salient morphological features; however, its Grad-CAM maps appear more diffused compared to deeper architectures, suggesting less discriminative spatial weighting. ShuffleNet-V2 displays broader activation regions, occasionally extending beyond the mushroom boundaries, indicating reduced spatial selectivity; nonetheless, Grad-CAM++ refines these maps by suppressing background activation. Overall, Grad-CAM++ consistently enhances fine-structure visibility across all models, with DenseNet121 and EfficientNet-B3 demonstrating the most biologically meaningful and class-discriminative focus patterns. Figures 29–30 present Grad-CAM and Grad-CAM++ maps for pairwise ensembles, highlighting the effect of feature combination on focus regions.

In the pairwise ensemble models, the Grad-CAM and Grad-CAM++ maps reveal how combined feature representations from two architectures influence the focus regions for each mushroom species. The DenseNet121 + EfficientNet-B3 ensemble (0.4, 0.6) produces well-localized attention across all three species, with EfficientNet's higher weight slightly sharpening class-specific regions. The DenseNet121 + MobileNet-V3 ensemble (0.5, 0.5) balances global contextual cues with finer local details, enhancing the stalk and cap margins in Amanita muscaria while preserving clear gill patterns in Mycena pura. DenseNet121 + ShuffleNet-V2 (0.5, 0.5) retains core structural focus but shows slightly

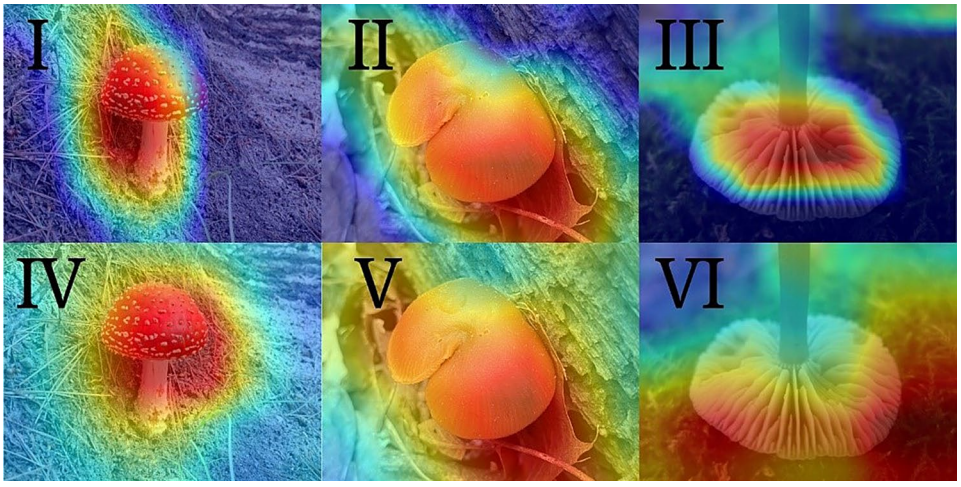

Fig. 29 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model Densenet121 + EfficientNet-B3, showing the attention focus for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

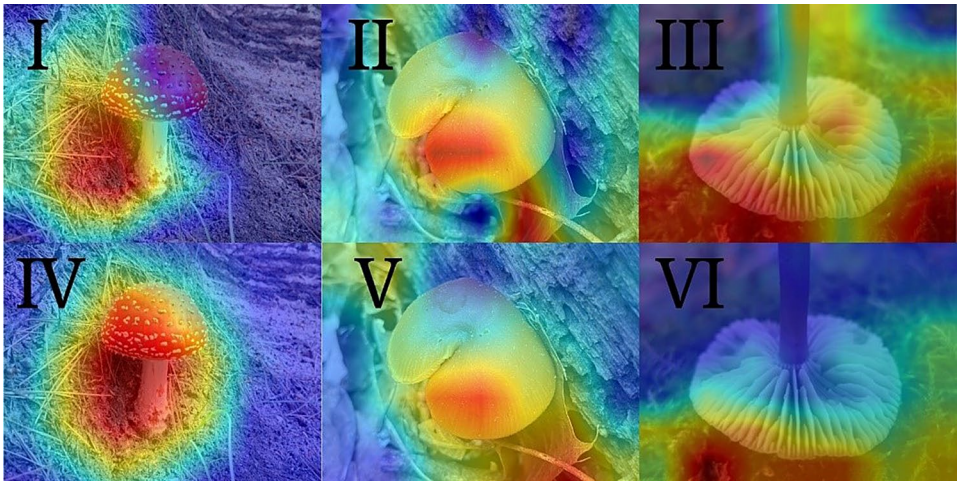

Fig. 30 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the pairwise ensemble model MobileNet-V3 + ShuffleNet-V2, showing the attention focus for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

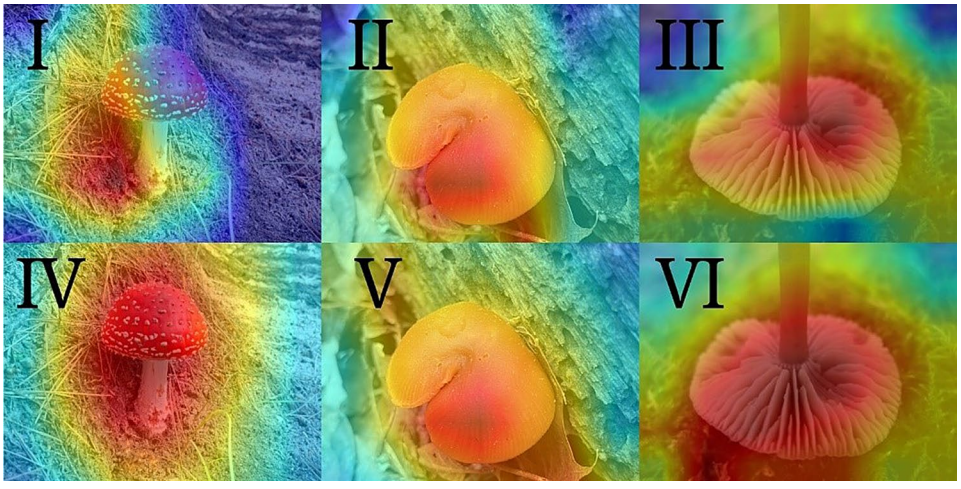

Fig. 31 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the triple ensemble model DenseNet121 + EfficientNet-B3 + ShuffleNet-V2, illustrating the key activation regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

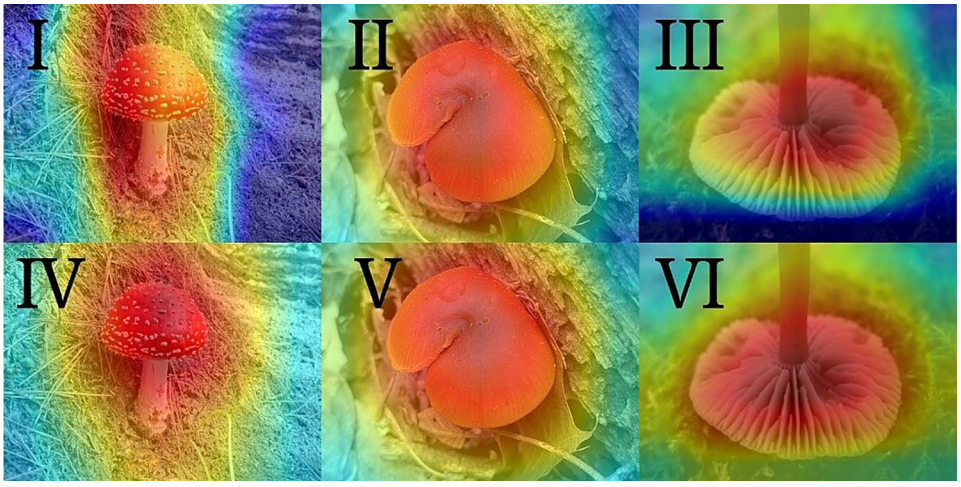

Fig. 32 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the triple ensemble model DenseNet121 + MobileNet-V3 + ShuffleNet-V2, illustrating the key activation regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

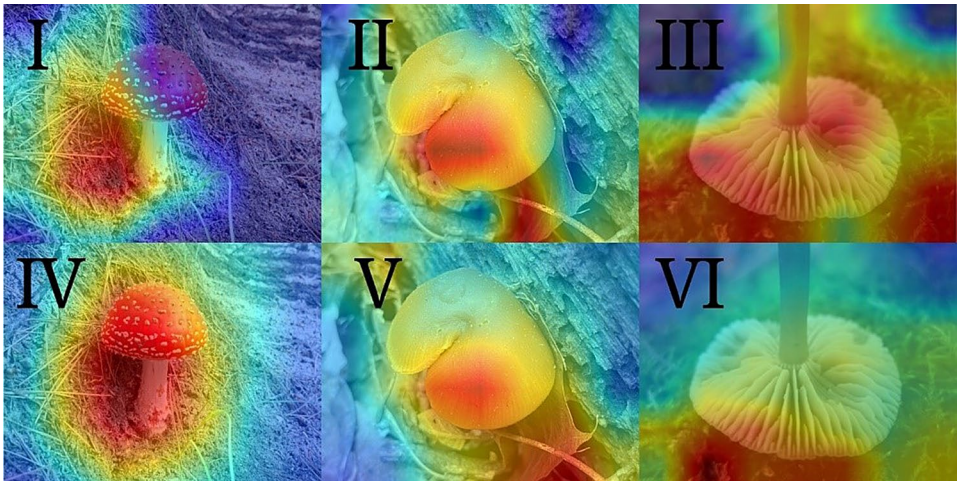

Fig. 33 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the triple ensemble model DenseNet121 + EfficientNet-B3 + MobileNet-V3, illustrating the key activation regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

broader heat dispersion, suggesting complementary but less sharply defined localization. The EfficientNet-B3 + MobileNet-V3 ensemble (0.5, 0.5) benefits from EfficientNet's strong high-level features and MobileNet's edge sensitivity, resulting in precise delineation of key morphological features. EfficientNet-B3 + ShuffleNet-V2 (0.7, 0.3) shows EfficientNet-dominant attention maps with strong central focus and reduced peripheral noise. Finally, the MobileNet-V3 + ShuffleNet-V2 ensemble (0.5, 0.5) captures both fine textural patterns and general shape cues, although the focus appears slightly diffused compared to more complex pairings. Overall, pairwise ensembles tend to combine the strengths of individual models, often leading to more stable and interpretable attention regions compared to their single-model counterparts. Figures 33–34 illustrate Grad-CAM and Grad-CAM++ maps for triple ensembles, showing enhanced focus on species-specific structures.

In the triple ensemble models, the Grad-CAM outputs generally focused on the distinctive regions of the target mushroom species, while Grad-CAM++ further refined these attention areas. In the DenseNet121 + EfficientNet-B3 + MobileNet-V3 combination (0.4, 0.5, 0.1), the high weight of EfficientNet-B3 emphasized morphologically critical regions (e.g., the cap and stipe in Amanita, pileus margin in Galerina, and gills in Mycena), with DenseNet121 contributing to the preservation of contextual integrity, and MobileNet-V3 due to its lower weight reducing peripheral noise. The DenseNet121 + EfficientNet-B3 + ShuffleNet-V2 combination (0.4, 0.4, 0.2) provided balanced class-specific focus, with ShuffleNet-V2's contribution mitigating background interference and enhancing the visibility of lamellar and pileus structures.

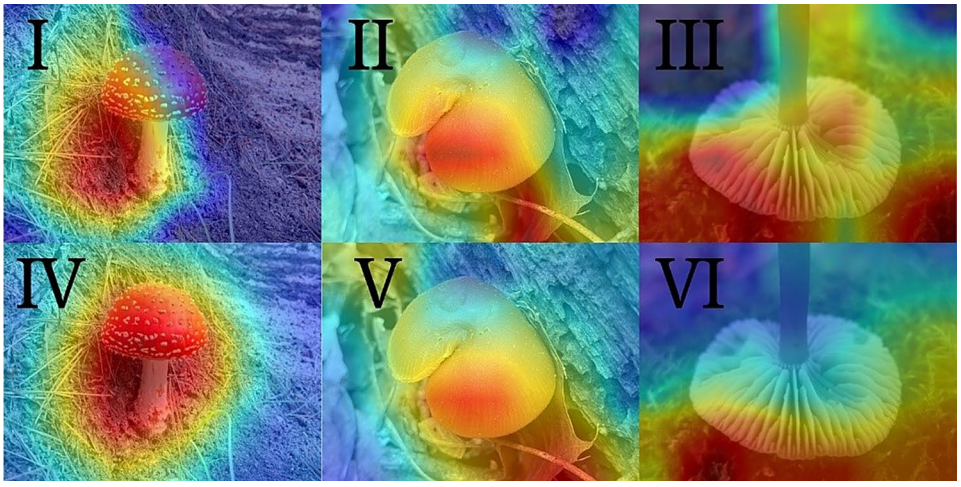

Fig. 34 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the triple ensemble model EfficientNet-B3 + MobileNet-V3 + ShuffleNet-V2, illustrating the key activation regions for *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

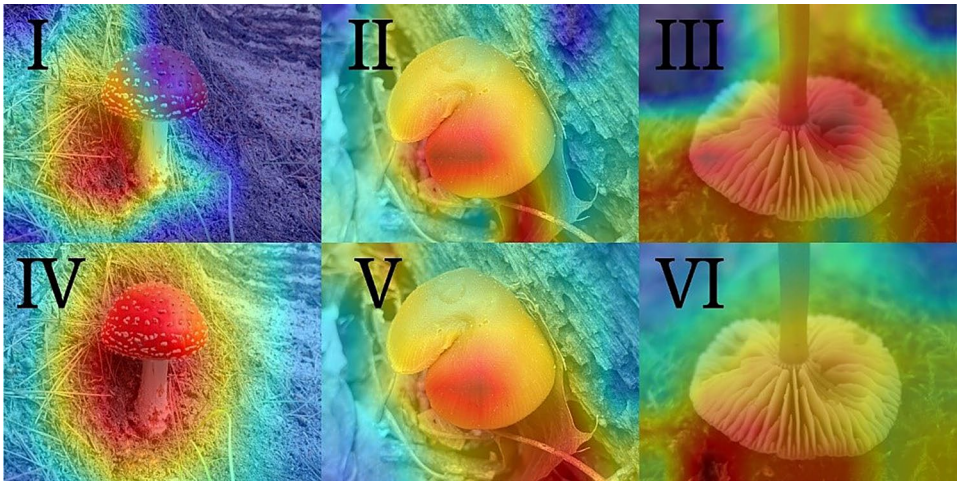

Fig. 35 Grad-CAM (top row) and Grad-CAM++ (bottom row) visualizations for the four-model ensemble DenseNet121 + EfficientNet-B3 + MobileNet-V3 + ShuflleNet-V2, depicting the salient areas contributing to the classification of *Amanita muscaria* (I, IV), *Galerina marginata* (II, V), and *Mycena pura* (III, VI)

In the DenseNet121 + MobileNet-V3 + ShuffleNet-V2 combination (0.4, 0.5, 0.1), MobileNet-V3's smooth and holistic emphasis on the fruiting body was shifted towards more specific anatomical regions (e.g., the volva–stipe junction in Amanita) through Grad-CAM++, while the low-weight ShuffleNet-V2 slightly expanded the receptive field. The EfficientNet-B3 + MobileNet-V3 + ShuffleNet-V2 combination (0.5, 0.1, 0.4) benefited from ShuffleNet-V2's higher contribution to increase textural awareness; although this occasionally expanded background attention in some species, the dominant EfficientNet-B3 weight ensured the preservation of class-relevant morphological cues, with Grad-CAM++ filtering out irrelevant activations. Figure 35 shows the Grad-CAM and Grad-CAM++ maps of the four-model ensemble, highlighting its balanced focus on global and detailed features.

In the four-model ensemble (DenseNet121 + EfficientNet-B3 + MobileNet-V3 + ShuffleNet-V2, 0.3, 0.4, 0.2, 0.1), the Grad-CAM outputs exhibited a harmonized integration of the individual models' attention mechanisms, effectively capturing both global contextual cues and fine-grained morphological details. The dominant contribution from EfficientNet-B3 (0.4) ensured consistent emphasis on key anatomical structures, such as the pileus and stipe in Amanita muscaria, the pileus curvature in Galerina marginata, and the lamellar architecture in Mycena pura. DenseNet121's influence (0.3) reinforced contextual coherence, reducing background distractions, while MobileNet-V3's moderate weight (0.2) added a smooth and contiguous focus over the fruiting body. ShuffleNet-V2's smaller contribution (0.1) provided localized texture sensitivity without significantly dispersing attention to irrelevant areas. Grad-CAM++ further

refined the highlighted regions, yielding sharper boundaries around diagnostic features, which collectively indicates that the four-model ensemble not only balances the strengths of its components but also mitigates their individual weaknesses in visual interpretability.

Across all individual and ensemble configurations, the Grad-CAM and Grad-CAM++ visualizations revealed a consistent focus on diagnostically relevant regions, particularly the pileus, lamellae, and stipe structures of the target mushroom species. Individual models demonstrated distinct attention patterns some emphasizing fine textural details while others favored broader contextual regions leading to varying strengths in feature localization. Pairwise and triple ensemble models exhibited more balanced heatmaps, effectively integrating complementary attention tendencies from their constituent networks, which enhanced both spatial coverage and feature precision. The four-model ensemble further unified these strengths, producing the most coherent and sharply defined activation patterns, with minimal background interference. Grad-CAM++ consistently provided more localized and refined boundaries compared to Grad-CAM, underscoring subtle morphological features critical for accurate classification. Overall, ensemble strategies proved advantageous in merging diverse model perspectives, yielding interpretability maps that were both more stable and semantically aligned with key morphological traits.

Discussion

Among the four CNN architectures compared in this study, EfficientNet-B3 demonstrated the highest performance. In terms of test accuracy, EfficientNet-B3 ranked first with 93.0% [41], followed by MobileNet-V3 (91.0%) [40], DenseNet121 (90.0%) [42], and ShuffleNet-V2 (86.78%) [43]. This indicates that EfficientNet-B3 achieved 3.0% higher accuracy than DenseNet121, 2.0% higher than MobileNet-V3, and 6.22% higher than ShuffleNet-V2. In terms of the Matthews Correlation Coefficient (MCC), EfficientNet-B3 reached the highest value at 92.60% [43], which is 3.15% higher than DenseNet121, 2.11% higher than MobileNet-V3, and 6.57% higher than ShuffleNet-V2. Furthermore, regarding log loss, EfficientNet-B3 (0.2549) produced approximately 39% lower error compared to DenseNet121 (0.4192), highlighting its ability to provide more reliable outputs. These differences clearly demonstrate that the compound scaling strategy of EfficientNet-B3 enables more successful representation of complex morphological variations.

The findings of this study are consistent with previous research in different disciplines. In agricultural image classification, EfficientNet-based models have been reported to achieve high success rates in the accurate detection of plant leaf diseases [20, 44]. Similarly, MobileNet-V3 has shown promising results in histopathological image analysis [45], offering a favorable balance between classification accuracy and rapid processing speed. When considered together, these results suggest that deeper and more scalable architectures (DenseNet121, EfficientNet-B3) provide stronger outcomes in biologically complex classification tasks [20], while lighter architectures (MobileNet-V3, ShuffleNet-V2) offer advantages in applications requiring high-speed performance [46]. Thus, the characteristics observed in this study are consistent with those reported across other fields, pointing to the cross-disciplinary generalizability of these models.

The mushroom dataset used in this study, with its balanced class distribution across 18 species and extensive data augmentation, offered strong generalization capacity. Having an equal number of samples per class provided a clear advantage over many health-related datasets, where class imbalance is frequently encountered. In ecological datasets, variability in background, lighting, and environmental conditions often presents challenges [47], yet the application of augmentations such as rotation, brightness/contrast adjustment, and noise injection in this study contributed to the model's adaptability to such diversity [48]. Moreover, the use of ensemble strategies was consistent with similar approaches reported in the literature (e.g., CNN–Transformer hybrid classification of puffball species, explainable AI-based classification of Discomycetes) [15]. In this context, the four-model ensemble not only achieved the highest accuracy (95.67%) [49] and MCC (95.42%) [50] but also demonstrated reliability in line with ensemble-based approaches reported across different disciplines.

In conventional mycological practice, mushroom species are typically identified based on a combination of macroscopic characteristics (cap morphology, color, gill structure, presence of volva or ring), microscopic features (spore size and ornamentation, basidia and cystidia structures), chemical reactions (KOH, Melzer's reagent), and ecological attributes (habitat, substrate, symbiotic plant species) [51, 52]. These methods, although effective, often require expert mycologists, are time-consuming, and may lead to misidentification when morphologically similar species are involved. The integration of deep learning methods in the current study therefore provides a complementary and efficient alternative to overcome these limitations [51–57].

Given the high overall accuracy levels (93–95.67%), the analysis of misclassified samples indicates that the remaining errors primarily arise from natural visual similarities between certain species and from environmental variability in the images. Most misclassifications occurred between morphologically similar species pairs, particularly *Mycena pura* – *Mycena rosea* and

Hypholoma fasciculare – *Galerina marginata*, which share overlapping macroscopic features under certain conditions. In addition, factors such as camera angle, lighting variation, and background diversity contributed to minor inconsistencies in classification. Therefore, incorporating more diverse imaging conditions and segmentation-based preprocessing steps may help further reduce these residual errors in future studies.

Conclusions

This study demonstrated that deep learning architectures can achieve highly accurate classification of wild poisonous mushroom species. Among the individual models, EfficientNet-B3 yielded the highest accuracy (93.0%), MCC (92.60), and the lowest log loss (0.2549). However, the most remarkable results were obtained with the four-model ensemble (DenseNet121, EfficientNet-B3, MobileNet-V3, ShuffleNet-V2), which achieved 95.67% accuracy, 95.42% MCC, and a log loss of 0.1751. Explainable AI analyses using Grad-CAM and Grad-CAM++ confirmed that the models consistently focused on biologically meaningful morphological features such as pileus, gills, and stipe. These results not only advance fungal taxonomy but also provide a reliable methodological framework with potential to reduce poisoning incidents caused by misidentification.

Beyond mushroom taxonomy, the proposed deep learning and ensemble approach can serve as a reference methodology for other domains where accurate visual classification is critical. In the medical field, it could be adapted for cancer detection and sub-type classification in histopathological images. In agriculture, the same framework can support the identification of crop diseases, monitoring of soil fertility, and assessment of plant growth conditions. In environmental sciences, it could be applied to evaluate water quality parameters and detect ecological anomalies. Importantly, in the context of food safety and consumer protection, this framework provides a practical foundation for digital quality control systems capable of detecting hazardous or substandard products before they reach the market. Such an integration into agricultural production pipelines would align with the broader vision of smart and sustainable agriculture. Therefore, this study not only contributes directly to the accurate classification of mushroom species but also illustrates the broader applicability of explainable ensemble-based deep learning methods across diverse health, agricultural, and environmental contexts.

Acknowledgements
We would like to thank BESTNAZTEK AI Systems Bilişim A.Ş. for providing the technical infra-structure that enabled us to obtain data for the artificial intelligence algorithms used in this study.

Author contributions
Conceptualization, A.A., H.S.K., F.E. and M.S.G.; methodology, A.F.K., M.S. and I.A.; software, A.A., H.S.K., F.E. and M.S.G.; validation, M.S.G., E.K., M.S. and F.E.; formal analysis, A.A., H.S.K., and I.A.; investigation, F.E. and M.S.G.; resources, E.K. and I.A.; writing—original draft preparation, A.A., H.S.K., M.S., I.A. and F.E.; writing—review and editing, M.S.G. and E.K.; visualization, A.A., H.S.K., A.F.K. and F.E.; supervision, M.S.G., I.A. and F.E. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Data availability
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethical approval
Not applicable.

Consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 October 2025 / Accepted: 18 December 2025
Published online: 25 December 2025

References

1. Tawatsin A, Parnmen S, Thavara U, Siriyasatien P, Kongtip P. Mushroom poisoning in Thailand: incidence and intoxication to human health. Med Res Archives. 2018;6(9).
2. Diaz JH. Amatoxin-containing mushroom poisonings: species, toxidromes, treatments, and outcomes. Wildern & Environ Med. 2018;29(1):111–18.
3. Gul H, Khan SM, Zeb SA, Maula F, Bibi N. A review on mysterious poisonous mushrooms of Pakistan, unveiling their implications and threats.
4. Yilmaz I, Ermis F, Akata I, Kaya E. A case study: what doses of Amanita phalloides and amatoxins are lethal to humans? Wildern & Environ Med. 2015;26(4):491–96.
5. Akata I, Yilmaz I, Kaya E, Coskun NC, Donmez M. Toxin components and toxicological importance of Galerina marginata from Turkey. Toxicon. 2020;187:29–34.
6. Singh K, Kumari A. Mushroom toxins and health hazards associated with them. In: Mycotoxins and mycotoxicoses. Springer. 2022. p. 105–27.
7. Lagrange E, Vernoux J-P. Warning on false or true morels and button mushrooms with potential toxicity linked to hydrazinic toxins: an update. Toxins. 2020;12(8):482.
8. Tkach VV, Morozova TV, Kushnir MV, de Oliveira SC, Kryvetskyi VV, Kryvetska II, Kryvetskyi IV, Biryuk IG, Sykyrytska TB, Yg I. Theoretical Description for psilocin and Coprine electrochemical determination in mushroom pulp and Biological liquids over cobalt (II) oxyhydroxide-modified electrode. 2024.
9. Didloff J, Boukes GJ, Nyambe MN, Beukes DR, Lerata MS, Vilane V, Lee M, Govender S, van de Venter M. Exploring antimycobacterial potential: safety evaluation and active compound isolation from Gymnopilus junonius. Antibiotics. 2025;14(2):179.
10. Stöver A, Haberl B, Helmreich C, Müller W, Musshoff F, Fels H, Graw M, Groth O. Fatal immunohaemolysis after the consumption of the poison pax mushroom: a focus on the diagnosis of the Paxillus syndrome with the aid of two case reports. Diagnostics. 2019;9(4):130.
11. White J, Weinstein SA, De Haro L, Bédry R, Schaper A, Rumack BH, Zilker T. Mushroom poisoning: a proposed new clinical classification. Toxicon. 2019;157:53–65.
12. Lima AD, Fortes RC, Novaes MG, Percario S. Poisonous mushrooms; a review of the most common intoxications. Nutricion Hospitalaria. 2012;27(2):402–08.
13. Ozsari S, Kumru E, Ekinci F, Akata I, Guzel MS, Acici K, Ozcan E, Asuroglu T. Deep learning-based classification of macrofungi: comparative analysis of advanced models for accurate fungi identification. Sensors. 2024;24(22):7189.

14. Ekinci F, Ugurlu G, Ozcan GS, Acici K, Asuroglu T, Kumru E, Guzel MS, Akata I. Classification of mycena and marasmius species using deep learning models: an ecological and taxonomic approach. Sensors. 2025;25(6):1642.
15. Korkmaz AF, Ekinci F, Altaş Ş, Kumru E, Güzel MS. Akata I: a deep learning and explainable AI-Based approach for the classification of Discomycetes species. Biology. 2025;14(6):719.
16. Kumru E, Ugurlu G, Sevindik M, Ekinci F, Güzel MS, Acici K, Akata I. Hybrid deep learning framework for high-accuracy classification of morphologically similar puffball species using CNN and Transformer architectures. Biology. 2025;14(7):816.
17. Kumru E, Korkmaz AF, Ekinci F, Aydoğan A, Güzel MS, Akata I. Deep ensemble learning and explainable AI for multi-class classification of earthstar fungal species. Biology. 2025;14(10):1313.
18. Kumru E, EKİNCİ F, Açici K, Altindal ÖB, Güzel MS. AKATA I: advanced deep learning approaches for the accurate classification of phallaceae fungi with explainable AI. Turk J Botany. 2025;49(5):388–405.
19. Trigka M, Dritsas E. A comprehensive survey of deep learning approaches in image processing. Sensors. 2025;25(2):531.
20. Ali H, Shifa N, Benlamri R, Farooque AA, Yaqub R. A fine tuned EfficientNet-B0 convolutional neural network for accurate and efficient classification of apple leaf diseases. Sci Rep. 2025;15(1):25732.
21. Huang J, Mei L, Long M, Liu Y, Sun W, Li X, Shen H, Zhou F, Ruan X, Wang D. Bm-net: Cnn-based mobilenet-v3 and bilinear structure for breast cancer detection in whole slide images. Bioengineering. 2022;9(6):261.
22. Zhou H, Su Y, Chen J, Li J, Ma L, Liu X, Lu S, Wu Q. Maize leaf disease recognition based on improved convolutional neural network ShuffleNetV2. Plants. 2024;13(12):1621.
23. Öztürk C, Taşyürek M, Türkdamar MU. Transfer learning and fine-tuned transfer learning methods' effectiveness analyse in the CNN-based deep learning models. Concurr Comput Pract Exp. 2023;35(4):e7542.
24. Amarù S, Marelli D, Ciocca G, Schettini R. Dalib: a curated repository of libraries for data augmentation in Computer Vision. J Imag. 2023;9(10):232.
25. Akhtar T, Gilani SO, Mushtaq Z, Arif S, Jamil M, Ayaz Y, Butt SI. Waris a: effective voting ensemble of homogenous ensembling with multiple attribute-selection approaches for improved identification of thyroid disorder. Electronics. 2021;10(23):3026.
26. Cerekci E, Alis D, Denizoglu N, Camurdan O, Seker ME, Ozer C, Hansu MY, Tanyel T, Oksuz I, Karaarslan E. Quantitative evaluation of saliency-based explainable artificial intelligence (XAI) methods in deep learning-based mammogram analysis. Eur J Radiol. 2024;173:111356.
27. Danilevicz MF, Upadhyaya SR, Batley J, Bennamoun M, Bayer PE, Edwards D. Understanding plant phenotypes in crop breeding through explainable AI. Plant Biotechnol J. 2025.
28. Diallo R, Edalo C, Awe OO. Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score. In: Practical statistical learning and data science methods: case studies from LISA, 2020 global network, USA. Springer; 2024. p. 283–312.
29. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.
30. Org OG. GBIF occurrence download. 2015.
31. Alomar K, Aysel HI, Cai X. Data augmentation in classification and segmentation: a survey and new strategies. J Imag. 2023;9(2):46.
32. Bello A, Ng S-C, Leung M-F. Skin cancer classification using fine-tuned transfer learning of DENSENET-121. Appl Sci. 2024;14(17):7707.
33. Kaur G, Sharma N, Gupta R. Wheat leaf disease classification using EfficientNet B3 pre-trained architecture. In: 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE). IEEE; 2023. p. 1-5.
34. Kumar Lilhore U, Simaiya S, Sharma YK, Kaswan KS, Rao KB, Rao VM, Baliyan A, Bijalwan A, Alroobaea R. A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization. Sci Rep. 2024;14(1):4299.
35. Chen Z, Yang J, Chen L, Jiao H. Garbage classification system based on improved ShuffleNet v2. Resour, Conserv Recycl. 2022;178:106090.
36. Kim S, Chae M, Lee J, Lee H. Advanced pharmaceutical recognition system based on deep learning for mobile medication identification. Appl Sci. 2025;15(10):5644.
37. Zhou L, Jain A, Dubey AK, Singh SK, Gupta N, Panwar A, Kumar S, Althaqafi TA, Arya V, Alhalabi W. FPA-based weighted average ensemble of deep learning models for classification of lung cancer using ct scan images. Sci Rep. 2025;15(1):19369.
38. Ren B, Guo H, Sun L, Wu Z, Timofte R, Li Y. The tenth NTIRE, 2025 efficient super-resolution challenge report. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025. p. 917–66.
39. Ennab M, Mcheick H. Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models. Mach Learn Knowl Extr. 2025;7(1):12.
40. Gao Y, Liu J, Li W, Hou M, Li Y, Zhao H. Augmented grad-cam++: super-resolution saliency maps for visual interpretation of deep neural network. Electronics. 2023;12(23):4846.
41. Si H, Wang Y, Zhao W, Wang M, Song J, Wan L, Song Z, Li Y, Fernando B, Sun C. Apple surface defect detection method based on weight comparison transfer learning with MobileNetV3. Agriculture. 2023;13(4):824.
42. Leung KH, Rowe SP, Pomper MG, Du Y. Comparison of CNN-based approaches for detection of COVID-19 on chest X-ray images. In: 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). IEEE; 2020. p. 1-3.
43. Zhang X, Feng Q, Zhu D, Liang X, Zhang J. Compressing recognition network of cotton disease with spot-adaptive knowledge distillation. Front Plant Sci. 2024;15:1433543.
44. Kotwal JG, Kashyap R, Shafi PM. Artificial driving based EfficientNet for automatic plant leaf disease classification. Multimedia Tools Appl. 2024;83(13):38209–40.
45. DeVoe K, Takahashi G, Tarshizi E, Sacker A. Evaluation of the precision and accuracy in the classification of breast histopathology images using the MobileNetV3 model. J Pathol Inf. 2024;15:100377.
46. Bu Y, Liu H, Li H, Murengami BG, Wang X, Chen X. Grading algorithm for orah sorting line based on improved ShuffleNet V2. Appl Sci. 2025;15(8):4483.
47. Pyke GH, Ehrlich PR. Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biol Rev. 2010;85(2):247–66.
48. Gong Y, Hou Y, Zhang C, Jiang M. Beyond augmentation: empowering model robustness under extreme capture environments. 2024 Int Joint Conf On Neural Networks (IJCNN): 2024. IEEE: 1-8.
49. Niu B, Zhuang X, He Y, Lin Z. Navigation deception signal detection based on ensemble learning. In: 2023 IEEE International Conference on Unmanned Systems (ICUS): 2023. IEEE; p. 859–64.
50. Niu Y, Li Z, Chen Z, Huang W, Tan J, Tian F, Yang T, Fan Y, Wei J, Mu J. Efficient screening of pharmacological broad-spectrum anti-cancer peptides utilizing advanced bidirectional encoder representation from transformers strategy. Heliyon. 2024;10(9).
51. Atri N, Kaur M, Sharma S. Characterization of lamellate mushrooms-an appraisal. In: Developments in fungal biology and applied mycology. Springer; 2017. p. 471–500.
52. Senanayake I, Rathnayaka A, Marasinghe D, Calabon M, Gentekaki E, Lee H, Hurdeal V, Pem D, Dissanayake L, Wijesinghe S. Morphological approaches in studying fungi: collection, examination, isolation, sporulation and preservation. Mycosphere. 2020;11(1):2678–754.
53. Panchbhai KG, Lanjewar MG, Naik AV. Modified MobileNet with leaky ReLU and LSTM with balancing technique to classify the soil types. Earth Sci Inf. 2025;18(1):77.
54. Panchbhai KG, Lanjewar MG, Malik C VV, P. Small size CNN (CAS-CNN), and modified MobileNetV2 (CAS-MODMOBNET) to identify cashew nut and fruit diseases. Multimedia Tools Appl. 2024;83(42):89871–91.
55. Panchbhai KG, Lanjewar MG. Enhancement of tea leaf diseases identification using modified SOTA models. Neural Comput Appl. 2025;37(4):2435–53.
56. Korkmaz AF, Ekinci F, Kumru E, Altaş Ş, Güneş SK, Yalçın AT, Güzel MS, Akata I. Combination ensemble and explainable deep learning framework for high-accuracy classification of wild edible Macrofungi. Biology. 2025;14(12):1644.
57. Kumru E, Ekinci F, Aydoğan A, Güzel MS, Sevindik M, Akata I. Explainable convolutional neural network architectures for high-performance taxonomic classification of gasteroid macrofungi. Sci Rep. 2025;15(1):40196.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.