**Received:** 23 July 2026; **Revised:** 01 November 2026; **Accepted:** 15 November 2026; **Published:** 20 November 2026
*Archs Sci. (2026) Volume 76, Issue 5 Pages 254-262, Paper ID 2026526.*
*https://doi.org/10.68304/as/76526*

# Comparative Analysis of YOLO-Inspired, EfficientNetB3, and Hybrid EfficientNetB3 Models for Animal Image Classification

**Abhishek Chettri[1,*], Anuj Rai[1], Nima Norbu Sherpa[1], Sumeet Rai[1], Rahul Shah[1] and Bindhya Rai[2]**
[1]Faculty of Science and Technology, The ICFAI University Sikkim, Gangtok, India
[2]Faculty of Liberal Arts, The ICFAI University Sikkim, Gangtok, India
Correspondence: abhishekchettriwtfmma@gmail.com

**Abstract** Animal image classification plays an important role in wildlife monitoring, biodiversity conservation, ecological analysis, and surveillance applications. Recent advances in deep learning have enabled automated recognition of animal species from images with improved classification performance. This study presents a comparative evaluation of three deep learning approaches for animal image classification: a YOLO-inspired multi-scale feature extraction model, an EfficientNetB3-based transfer learning model, and a hybrid EfficientNetB3 architecture integrating YOLO-inspired multi-scale feature extraction and an attention mechanism. Experiments were conducted using the Animals Detection Images Dataset obtained from Kaggle. To improve class representation and reduce severe class imbalance, animal classes containing fewer than 150 images were excluded, resulting in a filtered dataset of 24,592 images across 45 animal classes. The dataset was divided into 15,742 training images, 3,957 validation images, and 4,893 test images. All three models were evaluated using the same preprocessing pipeline and experimental settings based on accuracy, precision, recall, the F1-score, training accuracy, validation accuracy, training loss, and validation loss. The results show that EfficientNetB3 achieved the best overall classification performance, obtaining 91.00% Accuracy, 90.00% Precision, 91.00% Recall, and 90.00% F1-Score. The YOLO-inspired model achieved an Accuracy of 88.00%, while the hybrid EfficientNetB3 model achieved an Accuracy of 84.35%. Although the hybrid architecture integrates multi-scale feature extraction, transfer learning, and attention mechanisms, its increased architectural complexity did not translate into improved classification performance. The findings demonstrate that, under the experimental conditions considered in this study, EfficientNetB3 provides a more effective balance between classification performance and added architectural components for animal image classification.

**Index Terms** Animal Image Classification, EfficientNetB3, YOLO-Inspired Feature Extraction, Transfer Learning, Hybrid Deep Learning, Wildlife Recognition, Computer Vision.

## I. Introduction

Animal image classification has become an important research area in computer vision because of its applications in wildlife monitoring, biodiversity conservation, ecological analysis, smart surveillance, and automated animal recognition. Automated identification of animals from images can assist researchers and conservationists in monitoring wildlife populations and analyzing animal presence without requiring continuous manual observation. However, reliable animal classification remains challenging because animal appearance can vary considerably due to differences in pose, scale, illumination, background, and visual similarity among species [1], [2].

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved image classification and object recognition by enabling the automatic extraction of hierarchical visual features from raw images. CNN-based approaches have demonstrated strong performance across a wide range of computer vision applications [3], [4]. In parallel, YOLO-based architectures, introduced by Redmon and colleagues, have gained considerable attention because they process visual information efficiently for object detection and support feature representations at different spatial scales [5]–[7]. EfficientNet architectures, first developed by Tan and Le, have demonstrated strong image classification performance by using compound scaling to balance network depth, width, and input resolution [2], [8], [9]. These characteristics make both multi-scale feature extraction and EfficientNet-based transfer learning relevant approaches for animal image classification.

Several studies have investigated deep learning approaches for animal recognition and classification. Abuduweili et al. [2] demonstrated the value of pretrained image classification models, augmentation, and ensemble learning for camera-trap species recognition. Gomez-Donoso et al. [10] evaluated a multilevel wildlife classification system that included Effi-

cientNetB5 as a convolutional backbone and reported promising results for hierarchical species recognition. Similarly, Neeli et al. [9] demonstrated the applicability of EfficientNet for bird species recognition. Other studies have explored YOLO-based approaches for object detection and multi-scale feature representation, demonstrating their ability to process objects of different sizes and support efficient visual recognition [6], [7]. Hybrid deep learning approaches have also been investigated to combine complementary feature extraction and classification mechanisms [2], [11], [12].

Despite these developments, a direct and controlled comparison between YOLO-inspired multi-scale feature extraction, EfficientNetB3-based transfer learning, and a hybrid architecture combining both approaches remains insufficiently explored within the context of animal image classification. Existing studies commonly differ in datasets, preprocessing procedures, model configurations, and evaluation criteria, making it difficult to determine whether increased architectural complexity actually results in improved classification performance. In particular, the effectiveness of integrating YOLO-inspired multi-scale feature extraction and attention mechanisms with EfficientNetB3, compared with standalone EfficientNetB3 and YOLO-inspired approaches, requires systematic evaluation under identical experimental conditions.

To address this gap, this study presents a comparative evaluation of three deep learning approaches: (i) a YOLO-inspired multi-scale feature extraction model, (ii) an EfficientNetB3 classification model based on transfer learning, and (iii) a hybrid EfficientNetB3 model incorporating YOLO-inspired multi-scale feature extraction and an attention mechanism. All three approaches are compared using the same animal image dataset, preprocessing pipeline, and experimental settings to ensure a fair comparison. The resulting dataset contains 24,592 images belonging to 45 animal classes. Model performance is evaluated using accuracy, precision, recall, the F1-score, training accuracy, validation accuracy, training loss, and validation loss. The comparative analysis aims to determine whether combining multiple feature extraction mechanisms and attention improves animal image classification performance or whether the simpler transfer-learning-based EfficientNetB3 architecture provides a more effective balance between classification performance and added model components.

## II. Background and Literature Review

Animal image analysis has become an important application area of computer vision and deep learning, particularly for wildlife monitoring, biodiversity conservation, ecological analysis, and automated surveillance. Traditional image recognition approaches relied heavily on handcrafted features and conventional machine learning techniques, which often faced difficulties under variations in illumination, complex backgrounds, changes in animal pose, and visual similarities between species [1], [3]. The development of Convolutional Neural Networks (CNNs) addressed several of these limitations by enabling the automatic extraction of hierarchical features directly from image data [3], [4], [13]. Recent studies have consequently demonstrated the effectiveness of deep learning models for animal classification and object recognition tasks [2], [9], [14], [15].

### *A. Deep Learning Techniques for Animal Image Classification*

CNN-based architectures are widely used for image classification because they can automatically learn increasingly complex visual representations from input images. Chaganti et al. [16] compared CNN and Support Vector Machine (SVM) approaches for image classification and demonstrated the effectiveness of CNN-based learning for complex image recognition tasks. Jana et al. [15] investigated a modified CNN architecture for animal image classification and reported strong classification performance, demonstrating the applicability of CNNs to wildlife-related recognition tasks.

Fine-grained animal classification presents additional challenges because different species may exhibit highly similar visual characteristics. Zeng [17] applied a CNN-based approach to similar-animal classification and demonstrated its capability to distinguish visually similar animal categories. Similarly, Sharma et al. [4] compared logistic regression and CNN-based approaches for animal image classification and reported superior performance from the CNN model. These studies demonstrate the effectiveness of deep learning for learning discriminative visual representations without extensive dependence on manually engineered features.

### *B. YOLO-Based Object Detection and Multi-Scale Feature Representation*

Object detection is an important component of automated animal recognition because it enables both the localization and identification of objects within an image. YOLO (You Only Look Once) is a widely studied single-stage object detection architecture known for its relatively fast detection capability and suitability for real-time applications [6], [7], [18].

Mittal et al. [6] described YOLO as a single-stage regression-based object detection approach that performs detection efficiently. Viraktamath et al. [19] investigated YOLOv3 and demonstrated its capability to detect objects at different scales. Jiang et al. [18] reviewed the evolution of the YOLO architecture and highlighted improvements across successive versions in detection accuracy and computational efficiency. Wang and Liang [7] further examined the development of YOLO-based object detection approaches and discussed continued improvements in speed and detection performance.

Although YOLO is primarily designed for object detection, its multi-scale feature representations provide useful information for recognizing visual patterns at different spatial scales. Motivated by this characteristic, the present study employs a YOLO-inspired multi-scale feature extraction approach rather than treating a conventional YOLO detector as a direct image classification model. This distinction is important because the objective of the proposed approach is to investigate whether

multi-scale feature extraction inspired by YOLO can contribute effectively to animal image classification.

### C. EfficientNet-Based Image Classification

EfficientNet architectures were introduced to improve the balance between classification accuracy and computational efficiency through compound scaling of network depth, width, and input resolution [9], [12]. Rather than increasing a single network dimension independently, EfficientNet applies coordinated scaling to these dimensions, enabling the construction of models with improved accuracy while maintaining computational efficiency.

EfficientNet-based architectures have demonstrated promising results in animal and wildlife image classification. Abuduweili et al. [2] investigated transfer learning, augmentation, and ensemble learning for wildlife recognition, whereas Tan and Le [8] established EfficientNet's compound-scaling approach. Gomez-Donoso et al. [10] investigated multilevel wildlife classification using several pretrained backbones, including EfficientNetB5, across numerous categories. Neeli et al. [9] also demonstrated the applicability of EfficientNet for bird species recognition. These studies indicate that transfer learning with pretrained EfficientNet models can provide strong feature representations for animal image classification.

In the present study, EfficientNetB3 is employed as a transfer-learning-based classification model. The model is initialized using ImageNet-pretrained weights and subsequently adapted to the 45 animal classes considered in the experimental dataset. This provides a strong classification baseline against which the YOLO-inspired and hybrid approaches can be evaluated.

### D. Hybrid Deep Learning and Feature Fusion Approaches

Hybrid deep learning approaches attempt to combine complementary feature extraction or learning mechanisms within a single architecture. Such approaches can potentially provide richer feature representations by integrating information from multiple models or feature extraction stages [2], [11], [12].

Yin et al. [11] investigated a hybrid CNN-RNN architecture that combined image feature extraction with sequential dependency learning. Abuduweili et al. [2] explored a combination of object detection and image classification techniques for wildlife recognition. Hamdi et al. [12] investigated the fusion of pretrained CNN models for cat breed classification and demonstrated the potential of model fusion for improving classification performance.

However, combining multiple feature extraction mechanisms may also increase architectural complexity. Additional layers and feature representations can introduce redundant information, increase the number of trainable parameters, and make optimization more difficult. Therefore, increased architectural complexity does not necessarily guarantee improved classification performance. This motivates the experimental comparison of a standalone EfficientNetB3 model with a hybrid architecture incorporating additional YOLO-inspired multi-scale feature extraction and attention mechanisms.

### E. Research Gap, Comparative Framework, and Evaluation Measures

Previous research has demonstrated that CNN-based models are effective for hierarchical feature extraction and image classification [4], [15], [16]. EfficientNet-based architectures further improve the balance between classification performance and computational efficiency through compound scaling of network depth, width, and input resolution [2], [9], [10]. In contrast, YOLO architectures are primarily designed for fast object detection and provide effective multi-scale feature representations for objects of different sizes [6], [7], [18], [19]. Hybrid approaches combine multiple feature extraction or learning mechanisms and can potentially provide richer feature representations; however, the additional architectural components may also increase computational complexity, introduce redundant features, and make model optimization more difficult [2], [11], [12].

Despite these developments, direct comparisons among YOLO-inspired multi-scale feature extraction, EfficientNetB3-based transfer learning, and hybrid architectures remain limited in the context of animal image classification. Existing studies commonly employ different datasets, preprocessing procedures, hyperparameter configurations, and evaluation protocols, making it difficult to determine whether a more complex architecture consistently provides better classification performance. In particular, it remains necessary to investigate whether integrating YOLO-inspired multi-scale feature extraction and an attention mechanism with EfficientNetB3 provides an advantage over the standalone EfficientNetB3 architecture.

To address this gap, this study provides a controlled comparative evaluation of three approaches: a YOLO-inspired multi-scale feature extraction model, an EfficientNetB3 transfer-learning model, and a hybrid EfficientNetB3 model incorporating YOLO-inspired multi-scale feature extraction and attention. All three models are evaluated using the same animal image dataset, preprocessing pipeline, hyperparameter settings, training conditions, and evaluation criteria. This comparative experimental design supports a consistent assessment of the strengths and limitations of each approach and determines whether the additional architectural complexity of the hybrid model leads to improved classification performance.

The principal techniques considered in this study include CNN-based feature learning, YOLO-inspired multi-scale feature extraction, EfficientNetB3 transfer learning, and attention-based hybrid feature representation. The YOLO-inspired model uses multi-scale feature extraction to capture visual information at different spatial scales, while EfficientNetB3 uses ImageNet-pretrained representations for transfer-learning-based classification. The hybrid model combines EfficientNetB3 feature extraction with YOLO-inspired multi-scale processing and an attention mechanism to emphasize relevant feature representations.

Model performance is evaluated using accuracy, precision, recall, and the F1-score. In addition, training accuracy, validation accuracy, training loss, and validation loss are analyzed to

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

examine learning behavior, convergence, generalization, and potential overfitting or underfitting. The combined evaluation provides a multifaceted basis for comparing the classification performance and learning characteristics of the three approaches.

### *F. Evaluation Metrics and Literature Comparison*

Previous research has demonstrated that CNN-based models can learn hierarchical representations relevant to image recognition [4], [15], [16]. EfficientNet extends this approach through compound scaling of network depth, width, and input resolution [8]–[10]. These studies motivate the combined reporting of overall accuracy and class-sensitive measures rather than relying solely on an aggregate accuracy figure. A comparison based on several metrics is especially relevant to wildlife datasets, where visually similar animal categories and differences in class frequency may conceal disparities in model performance.

The performance of an animal image classification model should be evaluated using multiple complementary measures rather than relying on a single metric. Classification accuracy provides an overall measure of correctly predicted samples, while Precision, Recall, and F1-Score provide additional information about the correctness and completeness of class predictions. These metrics are particularly relevant for animal classification because the performance of a model may vary across different animal classes [2], [9], [10].

In this study, accuracy, precision, recall, and the F1-score are used as the primary performance metrics for comparing the YOLO-inspired, EfficientNetB3, and hybrid EfficientNetB3 models. Accuracy measures the proportion of correctly classified images among all evaluated images. Precision measures the proportion of correctly predicted samples among the samples predicted as a particular class, while Recall measures the proportion of correctly identified samples among all actual samples belonging to that class. The F1-score provides a combined measure of Precision and Recall and is useful when both types of classification error need to be considered.

In addition to the final classification metrics, training accuracy, validation accuracy, training loss, and validation loss are monitored during model training. These measures provide information about the learning behavior and generalization capability of the models. The relationship between training and validation performance can also help identify potential overfitting or underfitting during training [3], [4], [15].

The use of these metrics provides a multifaceted evaluation framework for comparing the three approaches in terms of classification performance, learning behavior, and generalization capability.

## III. Methodology and Architecture

This study presents a controlled comparative evaluation of three deep learning approaches for animal image classification: (i) a YOLO-inspired multi-scale feature extraction model, (ii) an EfficientNetB3-based transfer learning model, and (iii) a hybrid EfficientNetB3 model integrating YOLO-inspired multi-scale feature extraction and an attention mechanism. The overall workflow consists of dataset collection and preprocessing, model development, model training, and performance evaluation. Figure 1 illustrates the overall architecture and experimental workflow of the proposed framework.

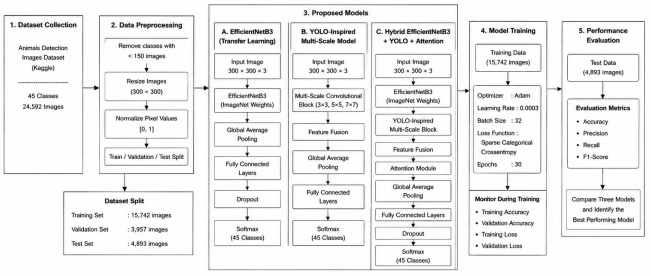



Figure 1: Overall workflow of the proposed animal image classification framework

### *A. Dataset Collection and Preprocessing*

The experiments were conducted using the Animals Detection Images Dataset available on Kaggle [20]. The source is an animal-detection collection containing image annotations; for the present image-level classification task, the dataset's class labels were used rather than training a bounding-box detector. The original dataset contains multiple animal classes with varying numbers of images per class. To reduce severe class imbalance and ensure sufficient samples for model training, classes containing fewer than 150 images were excluded from the dataset. After this filtering process, the resulting dataset contained 24,592 images belonging to 45 animal classes.

Table 1: Experimental setup and hyperparameter configuration

| Hyperparameter | Value |
|---|---|
| Dataset | Animal Detection Images Dataset (Kaggle) |
| Number of Classes | 45 |
| Total Images | 24,592 |
| Image Size | 300 × 300 |
| Batch Size | 32 |
| Optimizer | Adam |
| Learning Rate | 0.0003 |
| Loss Function | Sparse Categorical Crossentropy |
| Activation Function | Softmax |
| Epochs | 30 |
| Framework | TensorFlow / Keras |
| Transfer Learning | EfficientNetB3 (ImageNet Weights) |
| Training Images | 15,742 |
| Validation Images | 3,957 |
| Test Images | 4,893 |
| Validation Split | 20% |
| Programming Language | Python |
| Operating System | Windows 11 (Kaggle Notebook) |
| Hardware | Kaggle GPU |

The resulting dataset was divided into training, validation, and test subsets containing 15,742, 3,957, and 4,893 images, respectively. All input images were resized to 300 × 300 pixels to provide a consistent input dimension for the evaluated models. Pixel values were normalized before

being supplied to the models. A batch size of 32 was used during training. These reported subset counts sum to 24,592 images. The manuscript does not supply the random seed, exact per-class split assignments, image-duplication screening procedure, or augmentation rules; these details are therefore not reconstructed or assumed in the present report.

The same preprocessing procedure was applied to all three models to ensure a consistent comparison. The dataset composition and experimental configuration are summarized in Table 1.

### B. Experimental Setup

All three models were implemented using the TensorFlow/Keras framework and trained using the Adam optimizer with a learning rate of 0.0003 and a batch size of 32. The input images were resized to $300 \times 300$ pixels, and the models were trained for 30 epochs.

To ensure a fair comparison, the three approaches were evaluated using the same dataset partitions, preprocessing procedure, training duration, optimizer, learning rate, batch size, and classification objective. The EfficientNetB3-based models were initialized using ImageNet-pretrained weights. Model performance was evaluated on the independent test set after completion of training.

Using common partitions and reported training hyperparameters reduces some procedural differences between the experiments. However, the EfficientNetB3 models were initialized from pretrained weights, whereas the YOLO-inspired network used a different feature-learning strategy. Thus, the observed performance differences cannot be attributed solely to architectural complexity without further controls for initialization, parameter count, fine-tuning configuration, and class imbalance.

### C. EfficientNetB3-Based Classification Model

EfficientNetB3 was employed as the transfer-learning-based baseline classification model. EfficientNet uses a compound scaling strategy that jointly considers network depth, width, and input resolution to achieve an effective balance between model capacity and computational efficiency.

In this study, EfficientNetB3 was initialized using weights pretrained on the ImageNet dataset. The pretrained backbone provides general-purpose visual representations that can be adapted to the animal image classification task. Feature maps generated by the EfficientNetB3 backbone are subsequently processed by the classification layers.

The backbone was initially configured with frozen layers to preserve the pretrained representations, followed by fine-tuning of selected layers to adapt the learned features to the target animal classification task. The final classification layer uses a Softmax activation function to generate class probabilities for the 45 animal categories.

### D. YOLO-Inspired Multi-Scale Feature Extraction

The second approach uses a YOLO-inspired multi-scale feature extraction architecture. Unlike a conventional YOLO object detector, the proposed approach does not perform bounding-box-based object detection. Instead, it is inspired by the multi-scale feature representation used in YOLO-style architectures and applies multiple convolutional feature extraction branches to capture information at different spatial scales.

The multi-scale feature extraction block consists of convolutional operations with different kernel sizes, including $3 \times 3$, $5 \times 5$, and $7 \times 7$ convolutions. These parallel branches are intended to capture complementary visual information. Smaller kernels focus on local details, while larger kernels provide a broader contextual representation of animal appearance.

The resulting feature representations are combined through feature fusion and subsequently passed to the classification stage. This design enables the model to investigate whether multi-scale feature extraction alone can provide effective representations for animal image classification.

### E. Hybrid EfficientNetB3 Architecture with YOLO-Inspired Feature Extraction and Attention

The third approach combines the feature representation capability of EfficientNetB3 with YOLO-inspired multi-scale feature extraction and an attention mechanism. The objective of this architecture is to determine whether combining pretrained deep features, multi-scale representations, and attention can improve animal image classification performance.

For an input image of size $300 \times 300 \times 3$, EfficientNetB3 first extracts high-level feature representations. These feature maps are then passed to the YOLO-inspired multi-scale feature extraction block, where parallel convolutional branches with different kernel sizes capture complementary local and contextual information. The resulting feature maps are fused before being passed to the attention module.

The attention mechanism assigns different weights to the extracted feature representations, allowing the network to emphasize more informative features and reduce the influence of less relevant information. The attended features are subsequently processed through global average pooling and fully connected layers. A dropout layer is incorporated as a regularization mechanism before the final Softmax classification layer, which produces probability estimates for the 45 animal classes.

The hybrid architecture is represented within the overall workflow shown in Figure 1. The architecture follows the sequence of EfficientNetB3 feature extraction, YOLO-inspired multi-scale feature extraction, feature fusion, attention processing, global average pooling, fully connected classification, dropout regularization, and Softmax output.

The hybrid architecture introduces additional feature extraction and attention components compared with the standalone EfficientNetB3 model. Although these components may provide richer feature representations, they also increase architectural complexity and may introduce redundant features or optimization difficulties. Therefore, the hybrid architecture is evaluated experimentally to determine whether the

additional complexity provides a measurable improvement in classification performance.

### F. Model Training and Evaluation

All three models were evaluated under the reported common experimental conditions to support a consistent comparison, while their architectures and initializations differed. The dataset was divided into training, validation, and independent test subsets, and the same preprocessing procedure was applied to each model.

During training, training accuracy, validation accuracy, training loss, and validation loss were monitored to analyze the learning behavior and generalization capability of each model. After training, the final models were evaluated using the independent test set.

Classification performance was measured using accuracy, precision, recall, and the F1-score. These metrics provide complementary information about the overall classification performance and the correctness and completeness of model predictions.

The final comparison considers both classification performance and learning behavior. The available results do not report inference latency, parameter counts, confidence intervals, repeated-run variability, or class-level confusion matrices; consequently, performance differences should be interpreted within the reported evaluation rather than as a verified speed or computational-efficiency ranking. This allows the study to determine whether the additional multi-scale feature extraction and attention mechanisms used in the hybrid architecture provide an advantage over the standalone EfficientNetB3 and YOLO-inspired approaches.

Algorithm 1. Overall Experimental Workflow
*Input: Animal image dataset*
*Output: Predicted Animal Class*

Step 1: Load the Animals Detection Images Dataset from Kaggle.

Step 2: Exclude all animal classes containing fewer than 150 images.

Step 3: Resize all images to 300 × 300 pixels

Step 4: Normalize image pixel values to the range [0,1].

Step 5: Divide the dataset into training, validation, and independent test subsets.

Step 6: Initialize the EfficientNetB3 model using ImageNet pre-trained weights.

Step 7: Construct the YOLO-inspired multi-scale feature extraction block using multiple convolutional scales.

Step 8: Construct the hybrid architecture by integrating EfficientNetB3, YOLO-inspired multi-scale feature extraction, feature fusion, and attention.

Step 9: Train all three models using identical hyperparameter settings.

Step 10: Monitor training and validation performance during each training epoch.

Step 11: Evaluate each trained model using the independent test dataset.

Step 12: Compute accuracy, precision, recall, and the F1-score.

Step 13: Analyze training accuracy, validation accuracy, training loss, and validation loss.

Step 14: Compare the performance and learning behavior of the three models.

Step 15: Identify the best-performing approach based on the evaluation results.

*End*

## IV. Experimental Results and Discussion

### A. Evaluation Context

The experiments were conducted using the Animals Detection Images Dataset obtained from Kaggle. After excluding animal classes containing fewer than 150 images, the final dataset consisted of 24,592 images belonging to 45 animal classes. The three models were evaluated using the same dataset partitions, preprocessing procedure, and experimental settings to ensure a consistent comparison.

All models were implemented using Python with the TensorFlow/Keras framework. The EfficientNetB3-based models used ImageNet-pretrained weights for transfer learning. The models were trained using the Adam optimizer with a learning rate of 0.0003, a batch size of 32, and 30 training epochs. All input images were resized to 300 × 300 pixels.

Model performance was evaluated using accuracy, precision, recall, and the F1-score. training accuracy, validation accuracy, training loss, and validation loss were additionally analyzed to examine the learning behavior and generalization characteristics of the models.

### B. Performance Comparison

The results below compare the three models on the same held-out image set using four reported classification metrics. In addition to overall accuracy, precision, recall, and the F1-score help characterize the balance between false-positive and false-negative classifications. Because the manuscript does not identify the macro-, micro-, or weighted-averaging convention used across 45 classes, the precision, recall, and F1-score values should be interpreted as reported summary statistics pending confirmation of their aggregation method. Figure 2 presents the graphical accuracy comparison, while Table 2 provides the numerical results.

Table 2: Test-set classification performance of the three models (%)

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| YOLO-inspired | 88.00 | 87.00 | 88.00 | 87.00 |
| EfficientNetB3 | 91.00 | 90.00 | 91.00 | 90.00 |
| Hybrid | 84.35 | 84.00 | 84.00 | 84.00 |

Table 2 presents the comparative classification performance of the three evaluated models. EfficientNetB3 achieved the highest performance across all four reported metrics, with 91.00% Accuracy, 90.00% Precision, 91.00% Recall, and

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

90.00% F1-Score. These results indicate that the transfer-learning-based EfficientNetB3 model provided the most effective classification performance among the three evaluated approaches under the experimental conditions of this study.

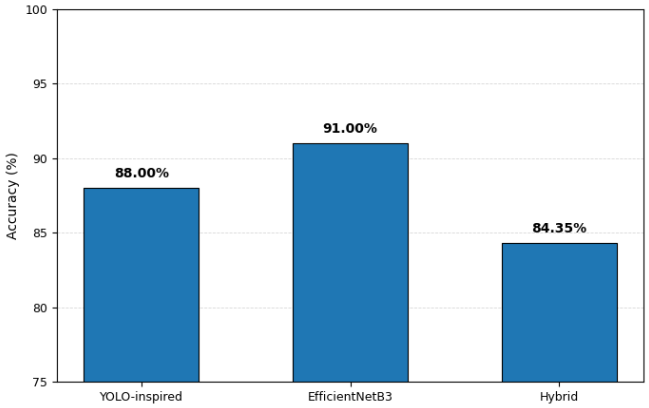



Figure 2: Accuracy comparison of the YOLO-inspired, EfficientNetB3, and hybrid EfficientNetB3 models

The YOLO-inspired model achieved 88.00% Accuracy, 87.00% Precision, 88.00% Recall, and 87.00% F1-Score. The results demonstrate that the multi-scale feature extraction approach was capable of learning useful visual representations for animal classification. However, its performance remained below that of EfficientNetB3, suggesting that the ImageNet-pretrained feature representations used by EfficientNetB3 provided an advantage for the classification task.

The hybrid EfficientNetB3 model achieved 84.35% Accuracy, 84.00% Precision, 84.00% Recall, and 84.00% F1-Score. Although the hybrid architecture combines EfficientNetB3, YOLO-inspired multi-scale feature extraction, and attention, the additional components did not result in improved classification performance under the evaluated experimental conditions. The lower performance suggests that the increased architectural complexity did not provide a sufficient classification advantage over the standalone EfficientNetB3 model.

Overall, the results show that the approach without the additional multi-scale fusion and attention components, EfficientNetB3, achieved the highest classification performance.

### C. Training and Validation Performance

To further examine the learning behavior of the three models, training accuracy, validation accuracy, training loss, and validation loss were analyzed. The corresponding results are presented in Table 3, while Figures 3 and 4 illustrate the training and validation performance.

EfficientNetB3 achieved the highest training and validation accuracy, with both values reaching 91.00%. The training and validation accuracy and loss values are identical at the reported precision. This pattern is compatible with similar aggregate performance across the splits, although rounded summary values alone cannot establish learning stability or rule out data leakage.

Table 3: Final training and validation performance across the evaluated models

| Model | Train Accuracy (%) | Validation Accuracy (%) | Train Loss | Validation Loss |
|---|---|---|---|---|
| YOLO-inspired Model | 88.00 | 88.00 | 0.4000 | 0.4000 |
| EfficientNetB3 | 91.00 | 91.00 | 0.3500 | 0.3500 |
| Hybrid EfficientNetB3 + YOLO-inspired | 88.58 | 86.18 | 0.2508 | 0.4620 |

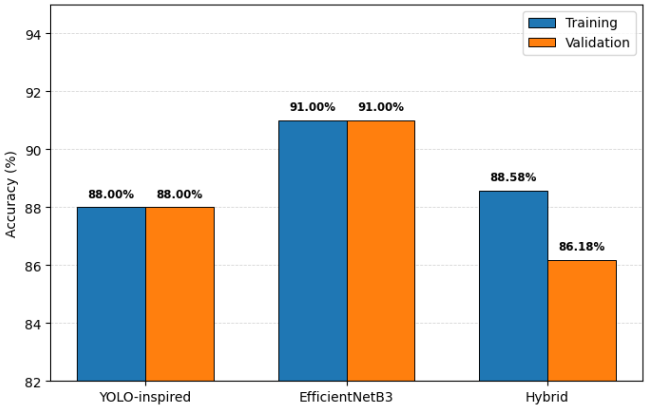



Figure 3: Training and validation accuracy of the three models

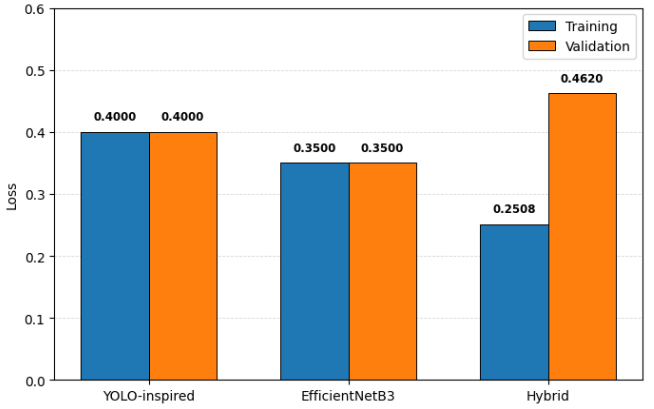



Figure 4: Training and validation loss of the three models

The YOLO-inspired model also demonstrated consistent training behavior, with both training and validation accuracy reported at 88.00% and both training and validation loss reported at 0.4000. The absence of a gap at the reported precision suggests similar aggregate scores for the two splits, but a fuller assessment of generalization would require the epoch-level histories and additional independent validation.

In contrast, the hybrid model achieved a training accuracy of 88.58% and a validation accuracy of 86.18%. Its training loss was 0.2508, while its validation loss increased to 0.4620. The difference between training and validation performance indicates a larger generalization gap than observed for the other two models. This behavior may indicate some degree of overfitting under the current experimental configuration.

The results therefore suggest that adding multi-scale feature extraction and attention to EfficientNetB3 did not improve the

learning or classification performance of the hybrid model. Instead, the standalone EfficientNetB3 model demonstrated more consistent training and validation behavior.

### D. Discussion

The experimental results demonstrate that EfficientNetB3 achieved the best overall classification performance among the three evaluated approaches. It obtained 91.00% Accuracy, 90.00% Precision, 91.00% Recall, and 90.00% F1-Score. The result indicates that the transfer-learning-based EfficientNetB3 architecture provided effective feature representations for the animal image classification task.

The YOLO-inspired model achieved 88.00% Accuracy, which was 3 percentage points lower than the Accuracy obtained by EfficientNetB3. Its Precision, Recall, and F1-Score were also lower by 3 percentage points. Nevertheless, the results demonstrate that multi-scale convolutional feature extraction can provide useful representations for animal classification. The lower performance compared with EfficientNetB3 may be associated with the absence of the pretrained ImageNet feature representation used by the EfficientNetB3-based approach.

The hybrid EfficientNetB3 model achieved the lowest classification performance, with an Accuracy of 84.35%, Precision of 84.00%, Recall of 84.00%, and F1-Score of 84.00%. The architecture was designed to combine the pretrained features of EfficientNetB3 with YOLO-inspired multi-scale feature extraction and attention. However, the additional architectural components did not improve the classification results under the current experimental conditions.

The training and validation results provide additional evidence for this observation. EfficientNetB3 achieved identical training and validation accuracy of 91.00%, while the YOLO-inspired model achieved identical values of 88.00%. In comparison, the hybrid model showed a training accuracy of 88.58% and validation accuracy of 86.18%, together with a training loss of 0.2508 and validation loss of 0.4620. The larger difference between training and validation loss suggests reduced generalization compared with the other two models.

These findings indicate that increasing architectural complexity does not necessarily lead to improved classification performance. In the present experiment, the additional YOLO-inspired feature extraction and attention components did not provide a measurable advantage over the standalone EfficientNetB3 model. Instead, EfficientNetB3 achieved the highest classification performance and the smallest reported difference between training and validation scores among the evaluated approaches.

Therefore, under the dataset and experimental configuration considered in this study, EfficientNetB3 represents the most effective of the three evaluated architectures for animal image classification in terms of classification performance and learning consistency.

## V. Conclusion

This study presented a comparative evaluation of three deep learning approaches for animal image classification: a YOLO-inspired multi-scale feature extraction model, an EfficientNetB3-based transfer learning model, and a hybrid EfficientNetB3 model integrating YOLO-inspired multi-scale feature extraction and an attention mechanism. All three models were evaluated using the same dataset partitions, reported preprocessing procedures, nominal training hyperparameters, and evaluation criteria to provide a consistent comparison, while model architecture and pretraining differed.

Among the evaluated approaches, EfficientNetB3 achieved the best overall classification performance, with 91.00% Accuracy, 90.00% Precision, 91.00% Recall, and 90.00% F1-Score. The YOLO-inspired model achieved 88.00% Accuracy, 87.00% Precision, 88.00% Recall, and 87.00% F1-Score, demonstrating that multi-scale feature extraction can provide useful representations for animal image classification. The hybrid EfficientNetB3 model achieved 84.35% Accuracy, 84.00% Precision, 84.00% Recall, and 84.00% F1-Score. Although the hybrid architecture integrates EfficientNetB3, YOLO-inspired multi-scale feature extraction, and attention mechanisms, the additional components did not improve classification performance under the experimental conditions considered in this study. The training and validation results also indicated a larger generalization gap for the hybrid model than for the standalone EfficientNetB3 and YOLO-inspired models.

Overall, the findings demonstrate that increasing architectural complexity does not necessarily result in improved classification performance. Under the dataset and experimental configuration used in this study, EfficientNetB3 provided the strongest classification performance and the smallest reported train–validation difference among the evaluated approaches. These findings highlight the importance of evaluating additional architectural components based on their empirical contribution rather than assuming that increased feature extraction complexity will automatically improve classification performance. Because the available results are single summary reports, future evaluations should specify the class-averaging convention, reproducible training seeds, model parameter counts, inference time, image-level leakage checks, and repeated-run uncertainty, so that the ranking of architectures can be assessed more rigorously.

## Acknowledgment

The authors would like to thank Mr. Rahul Shah for his valuable guidance and support throughout this research work. The authors also acknowledge and appreciate the academic support and facilities provided by The ICFAI University, Sikkim for successfully completing the study task.

## Author Contributions

All authors contributed equally to the conception and design of the study, data collection and analysis, and the drafting and

critical revision of the manuscript. All authors have read and approved the final version of the manuscript.

## Conflict of Interest

The authors declare no conflicts of interest regarding this manuscript.

## Funding Information

The authors received no financial support for this article.

## Data Availability

The study used the publicly available Animals Detection Images Dataset hosted on Kaggle [20].

## References

[1] Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*(Suppl.), 1199–1204. https://doi.org/10.1038/81479

[2] Abuduweili, A., Wu, X., & Tao, X. (2019). *Efficient method for categorize animals in the wild* [Preprint]. arXiv. https://arxiv.org/abs/1907.13037

[3] Elngar, A. A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., & Fawzy, N. (2021). Image classification based on CNN: A survey. *Journal of Cybersecurity and Information Management, 6*(1), 18–50. https://doi.org/10.54216/JCIM.060102

[4] Sharma, R. R., Khan, J., Sungheetha, A., Mahapatra, S., & Ghantasala, G. S. P. (2025). A dual-model approach to animal image classification: Logistic regression and CNN. In *Proceedings of the International Conference on Advances and Applications in Artificial Intelligence (ICAAAI 2025)* (pp. 208–217). Atlantis Press. https://doi.org/10.2991/978-94-6463-738-0_17

[5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

[6] Mittal, N., Vaidya, A., & Kapoor, S. (2019). Object detection and classification using YOLO. *International Journal of Scientific Research & Engineering Trends, 5*(2), 562–565. https://ijsret.com/wp-content/uploads/2019/05/IJSRET_V5_issue2_219.pdf

[7] Wang, C., & Liang, G. (2023). Overview of research on object detection based on YOLO. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering* (pp. 257–265). ACM. https://doi.org/10.1145/3652628.3652671

[8] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[9] Neeli, S., Guruguri, C. S. R., Kammara, A. R. A., Annepu, V., Bagadi, K., & Chirra, V. R. R. (2023). Bird species detection using CNN and EfficientNet-B0. In *2023 International Conference on Next Generation Electronics (NEleX)* (pp. 1–6). IEEE. https://doi.org/10.1109/NEleX59773.2023.10420966

[10] Gomez-Donoso, F., Escalona, F., Pérez-Esteve, F., & Cazorla, M. (2021). Accurate multilevel classification for wildlife images. *Computational Intelligence and Neuroscience, 2021*, Article 6690590. https://doi.org/10.1155/2021/6690590

[11] Yin, Q., Zhang, R., & Shao, X. (2019). CNN and RNN mixed model for image classification. *MATEC Web of Conferences, 277*, Article 02001.

[12] Hamdi, E. B., Sunaryo, J. A., & Prasetyo, S. Y. (2023). Fusion of pre-trained CNN models for cat breed classification: A comparative study. *E3S Web of Conferences, 426*, Article 01014. https://doi.org/10.1051/e3sconf/202342601014

[13] Lei, X., Pan, H., & Huang, X. (2019). A dilated CNN model for image classification. *IEEE Access, 7*, 124087–124095. https://doi.org/10.1109/ACCESS.2019.2927169

[14] Singh, A., Pietrasik, M., Natha, G., Ghouaiel, N., Brizel, K., & Ray, N. (2020). Animal detection in man-made environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1427–1438). IEEE. https://doi.org/10.1109/WACV45572.2020.9093504

[15] Jana, N. K., Chowdhury, A., Saha, D., Dutta, C., & Maiti, A. (2022). Classification of animal images with modified CNN model. *12th Inter-University Engineering, Science & Technology Academic Meet (FOSET 2022)*. https://doi.org/10.36375/prepare_u.foset.a299

[16] Chaganti, S. Y., Nanda, I., Pandi, K. R., Prudhvith, T. G. N. R. S. N., & Kumar, N. (2020). Image classification using SVM and CNN. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCSEA49143.2020.9132851

[17] Zeng, P. (2021). Research on similar animal classification based on CNN algorithm. *Journal of Physics: Conference Series, 2132*(1), Article 012001. https://doi.org/10.1088/1742-6596/2132/1/012001

[18] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of YOLO algorithm developments. *Procedia Computer Science, 199*, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135

[19] Viraktamath, S. V., Yavagal, M., & Byahatti, R. (2021). Object detection and classification using YOLOv3. *International Journal of Engineering Research & Technology, 10*(2), 197–202. https://www.ijert.org/object-detection-and-classification-using-yolov3

[20] antoreepjana. (n.d.). *Animals detection images dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/antoreepjana/animals-detection-images-dataset