



A HYBRID RESNET–INCEPTION DEEP LEARNING ARCHITECTURE FOR ENHANCED WILDLIFE IDENTIFICATION

Malach Obisa¹, Edna Too², Benard Osero³

^{1,2,3}Department of Computer Science, Chuka University, P.O.

Box 109 – 60400, Chuka, Kenya.

E-Mail: malachobisa@gmail.com

*Corresponding author

malachobisa@gmail.com

Article history:

Received June 5, 2026

Revised July 12, 2026

Accepted August 5, 2026

Keywords:

Hybrid Deep Learning,
Wildlife Identification,
Feature Fusion,
Computer Vision, Image
Classification, Ecological
Monitoring.

Abstract

Accurate wildlife identification in unconstrained environments remains a significant challenge due to intra-class similarity, occlusions, illumination variations, and complex backgrounds. This study of hybrid deep learning architecture that enhances classification performance by integrating complementary feature learning mechanisms within a unified framework. The model captures both deep hierarchical representations and multi-scale spatial features through a feature-level fusion strategy, resulting in a more discriminative and robust representation of wildlife images. The hybrid approach was trained and evaluated on the Animals with Attributes 2 (AwA2) dataset and further validated on the WildlifeReID-10k dataset to assess its generalization capability. Experimental results show that the model achieves a classification accuracy of 98%, along with high precision, recall, and F1-score, demonstrating strong reliability and effectiveness. Furthermore, the model maintains high performance on external data, confirming its robustness and ability to generalize across diverse real-world conditions. These findings highlight the effectiveness of hybrid deep learning approaches in addressing complex wildlife identification challenges and support their application in ecological monitoring and biodiversity conservation.

1.0 INTRODUCTION

Automated wildlife identification has become an essential component in modern biodiversity conservation, ecological monitoring, and anti-poaching efforts. With the rapid growth of camera trap deployments, drone surveillance, and large-scale ecological datasets, there is an increasing demand for intelligent systems capable of accurately identifying wildlife species from images. Traditional identification methods, which rely heavily on human experts, are not only time-consuming and labor-intensive but also prone to subjectivity and inconsistency. These limitations make them unsuitable for large-scale, real-time applications [1][2]. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly transformed image classification tasks. CNN-based models have demonstrated remarkable success in extracting complex visual features and achieving high accuracy across various domains, including medical imaging, facial recognition, and wildlife monitoring. Among the most influential architectures are Residual Networks (ResNet) and Inception networks, both of which have introduced innovative mechanisms to improve feature learning.

ResNet-101 addresses the degradation problem in deep neural networks through residual learning, enabling the training of very deep architectures. This allows the model to learn rich hierarchical representations that are crucial for distinguishing between visually similar classes. However, despite its depth, ResNet primarily focuses on sequential feature extraction and may not effectively capture fine-grained spatial variations that are critical in wildlife identification, where subtle differences in texture, shape, or patterns distinguish species [3]. On the other hand, Inception v3 employs parallel convolutional filters of varying sizes, allowing it to capture multi-scale spatial features within an image. This makes it particularly effective in handling variations in object size, orientation, and background complexity. Nevertheless, Inception networks are relatively shallower compared to ResNet and may lack the depth required to fully capture high-level semantic features necessary for robust classification.

These complementary strengths and limitations highlight a critical research gap: while individual architectures perform well in isolation, they may fail to fully address the complexity of real-world wildlife datasets. This challenge forms the foundation of this study, which focuses on the design and development of a hybrid deep learning architecture that integrates ResNet-101 and Inception v3. The hybrid model aimed to combine the deep hierarchical feature learning capability of ResNet with the multi-scale spatial feature extraction of Inception within a unified framework. By processing input images through parallel branches and fusing the extracted features, the model is expected to generate a more comprehensive and discriminative representation of wildlife images. This integration is hypothesized to improve classification accuracy, enhance generalization, and reduce misclassification, particularly in challenging scenarios involving occlusion, background clutter, and inter-class similarity [4].

Therefore, this study is strictly focused on the development, implementation, and evaluation of the hybrid ResNet–Inception model as defined in Objective 2. The emphasis is placed on architectural design, feature fusion strategy, and performance evaluation, with the aim of demonstrating the effectiveness of hybrid deep learning approaches in complex image classification tasks.

2.0 THEORETICAL

The field of image classification has experienced significant advancement with the emergence of deep learning, particularly Convolutional Neural Networks (CNNs). Over the past decade, several architectures have been proposed to improve feature extraction, model depth, and classification accuracy. This section reviews key developments relevant to this study, with a particular focus on architectures that motivate the design of the hybrid ResNet–Inception model.

2.1 Convolutional Neural Networks in Image Classification

CNNs have become the dominant approach for image-based tasks due to their ability to automatically learn hierarchical feature representations. Early architectures such as AlexNet and VGGNet demonstrated the effectiveness of deep convolutional layers in extracting visual features. However, these models suffered from limitations such as vanishing gradients and computational inefficiency as depth increased [5][6], [7], [8].

2.2 Residual Networks (ResNet)

Residual Networks introduced a breakthrough in deep learning by addressing the degradation problem associated with very deep networks. The key innovation in ResNet is the use of residual (skip) connections, which allow gradients to flow directly through the network, enabling the training of very deep architectures such as ResNet-50, ResNet-101, and beyond [9]. ResNet-101, in particular, is capable of learning highly abstract and hierarchical representations. This makes it suitable for distinguishing complex patterns in images. However, despite its depth, ResNet primarily processes features sequentially and may not effectively capture fine-grained spatial variations across different scales, which are essential in wildlife identification tasks [10].

2.3 Inception Networks

The Inception architecture introduced a different approach to feature extraction by using parallel convolutional filters of varying sizes within the same layer. This design enables the

network to capture features at multiple spatial scales simultaneously. Inception v3 further improved efficiency by incorporating factorized convolutions and dimensionality reduction techniques [11]. Inception networks are particularly effective in handling images with varying object sizes and complex backgrounds. However, compared to deeper networks like ResNet, they may lack sufficient depth to fully capture high-level semantic representations, which can limit their performance in complex classification tasks [12].

2.4 Hybrid Deep Learning Architectures

To overcome the limitations of individual architectures, recent research has explored hybrid models that combine multiple CNN architectures. These models aim to leverage the complementary strengths of different networks to improve feature representation and classification performance [13]. Hybrid approaches typically involve combining feature maps from different models through techniques such as concatenation, averaging, or attention-based fusion. Studies have shown that such approaches can significantly improve performance in domains such as medical imaging, object detection, and fine-grained classification [14]. However, there is still limited research focusing specifically on hybrid architectures for wildlife identification. This gap highlights the need for models that can effectively handle the complexity and variability present in wildlife datasets [14].

2.5 Summary of Related Work

The literature indicates that while ResNet excels in deep feature learning and Inception excels in multi-scale feature extraction, neither architecture alone is sufficient to address all challenges in wildlife identification. Hybrid architectures provide a promising solution by integrating complementary capabilities.

Table 1: Summary of Key Architectures and Their Characteristics

Model	Reference	Key Strength	Limitation	Relevance to This Study
AlexNet	[15]	Deep feature learning	Limited depth, overfitting	Foundation of CNNs
VGGNet	[16]	Simplicity and uniform architecture	High computational cost	Baseline reference
ResNet-101	[3]	Deep hierarchical feature extraction	Limited multi-scale sensitivity	Used as hybrid component
Inception v3	[17]	Multi-scale feature extraction	Limited depth	Used as hybrid component
Hybrid CNN Models	[18]	Combines strengths of multiple models	Increased complexity	Core focus of this study

2.6 Research Gap

Based on the reviewed literature, a clear gap exists in the development of hybrid architectures specifically tailored for wildlife identification. Most existing models focus on either depth or multi-scale representation, but not both simultaneously. This study addresses this gap by proposing a hybrid ResNet–Inception model that integrates deep hierarchical and multi-scale feature learning within a unified framework.

3.0 METHODOLOGY

This section presents a detailed and coherent description of the research design, focusing strictly on Objective 2, which involves the development and evaluation of a hybrid ResNet–Inception model. The methodology is structured to provide a smooth conceptual and procedural flow, enabling the reader to clearly understand how the research progresses from data preparation to model evaluation.

3.1 System Workflow and Research Design

The overall system follows a structured pipeline in which raw wildlife images are first collected and prepared before being processed through a hybrid deep learning architecture.

The workflow begins with dataset selection and preprocessing, where images are standardized to ensure compatibility with deep learning models. These processed images are then simultaneously fed into two parallel convolutional neural network branches—ResNet-101 and Inception v3. Within this architecture, each branch performs independent feature extraction. The ResNet branch focuses on capturing deep hierarchical representations, while the Inception branch captures multi-scale spatial features. The outputs from both branches are then combined at a feature fusion stage, producing a unified representation that is subsequently passed through fully connected layers for classification. Finally, the model is evaluated using both internal validation metrics and external datasets to assess its generalization capability.

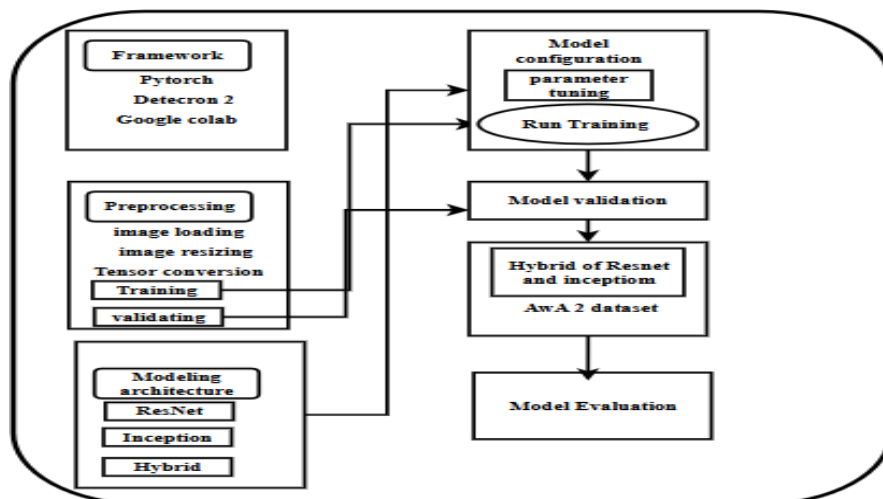


Figure 1: Conceptual architecture

The workflow diagram illustrates the sequential transformation of input images into classification outputs through preprocessing, dual-branch feature extraction, feature fusion, and final prediction. The workflow begins with the establishment of the development framework, where tools such as PyTorch, Detectron2, and Google Colab are utilized to support model implementation and experimentation. Once the environment is set up, the process proceeds to data preprocessing, where input images are loaded, resized to a consistent format, and converted into tensor representations suitable for deep learning models. Following preprocessing, the modeling stage is introduced, in which different architectures, namely ResNet, Inception, and the hybrid model—are defined and prepared for training. The next phase involves model training, where parameter tuning is conducted to optimize performance, followed by validation to assess the model's learning behavior and generalization capability. The hybrid model, trained using the AwA2 dataset, is then subjected to further validation to ensure robustness. Finally, the workflow concludes with model evaluation, where performance metrics are computed to determine the effectiveness and accuracy of the hybrid architecture in wildlife identification tasks.

3.2 Dataset Preparation and Preprocessing

The study utilizes the Animals with Attributes 2 (AwA2) dataset, which contains a wide range of wildlife images exhibiting variations in pose, lighting, and background conditions. To ensure computational efficiency and balanced representation, a subset of eleven classes was selected. Each image undergoes a series of preprocessing steps designed to improve model performance and training stability.

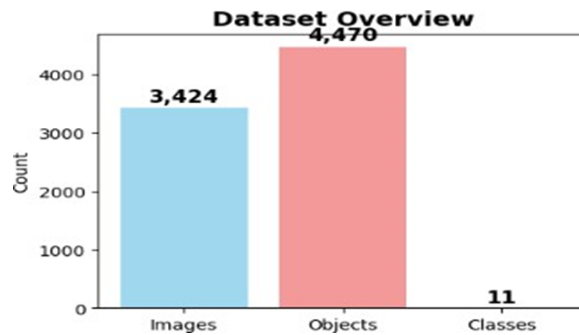


Figure 2: Dataset Preparation and Preprocessing

Initially, all images are resized to a uniform dimension to ensure compatibility with convolutional layers. This is followed by normalization, where pixel intensity values are scaled to a standardized range, thereby facilitating faster convergence during training. Finally, the images are converted into tensor representations, making them suitable for processing within the deep learning framework.

Table 2: Dataset Summary

Parameter	Description
Dataset	AwA2
Number of Classes	11
Data Type	RGB Images
Preprocessing	Resizing, Normalization, Tensor Conversion

3.3 Hybrid Model Architecture Design

The central contribution of this study lies in the design of a hybrid deep learning architecture that integrates ResNet-101 and Inception v3 within a unified framework. Unlike traditional sequential models, the hybrid architecture adopts a parallel structure, allowing both networks to process the same input image simultaneously. In this design, the ResNet-101 branch leverages residual connections to learn deep hierarchical features. These features are essential for capturing high-level semantic information that distinguishes different wildlife species. Meanwhile, the Inception v3 branch applies parallel convolutional filters of varying sizes, enabling the model to extract spatial features at multiple scales. This is particularly useful in handling variations in object size, orientation, and background complexity. The outputs of these two branches are not used independently; instead, they are combined at a fusion layer. This fusion step is critical, as it integrates the strengths of both architectures into a single, more expressive feature representation.

3.4 Data Distribution

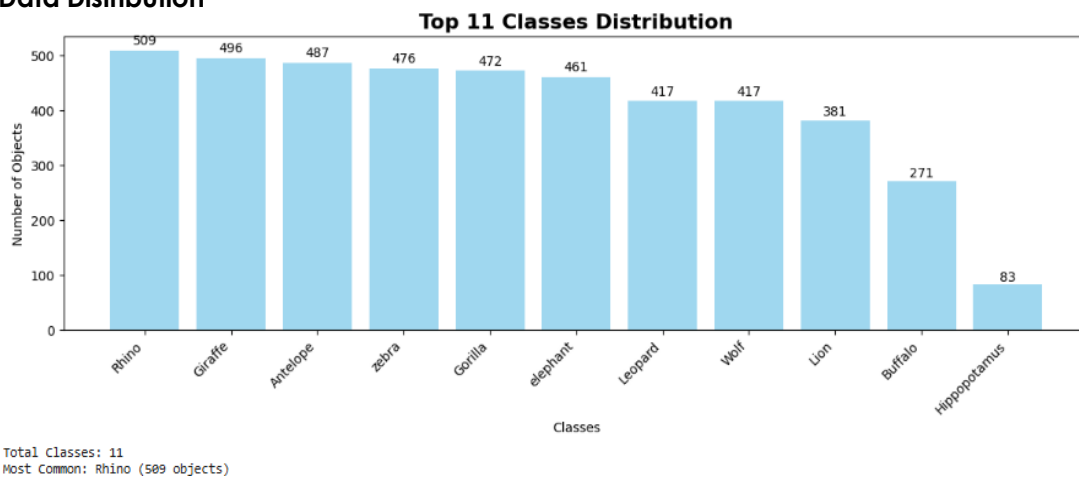


Figure 3: Class distribution for the 11 classes

The class distribution chart above illustrates the representation of the 11 wildlife species used in this study. Overall, the dataset demonstrates a relatively balanced distribution across most classes, with the number of objects per class ranging between 417 and 509 for the majority of categories. Notably, classes such as Rhino (509), Giraffe (496), Antelope (487), and Zebra (476) are well represented. Although a few classes like Buffalo (271) and Hippopotamus (83) have fewer instances, the dataset still maintains reasonable class diversity suitable for effective model training. This distribution supports robust learning and evaluation of the models without significant class dominance, ensuring fair representation across different species during classification.

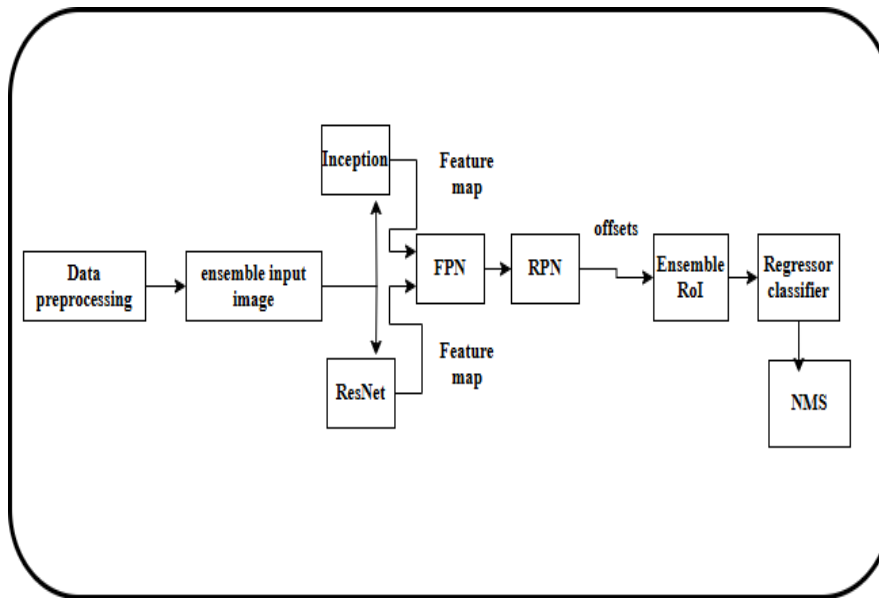


Figure 4: Hybrid ResNet-Inception Architecture

The architecture diagram shows how a single input image is processed through two parallel branches before their feature maps are merged and passed to classification layers.

3.5 Feature Extraction and Fusion Mechanism

Feature extraction within the hybrid model is performed independently by each branch. The ResNet-101 network produces a feature representation that emphasizes deep semantic information, while the Inception v3 network generates features that capture spatial variations across different scales. These feature representations are mathematically expressed as outputs from their respective networks and are subsequently combined through a concatenation operation. This fusion process results in a comprehensive feature vector that contains both hierarchical and spatial information. The fused features are then passed through fully connected layers, where they are transformed into class probabilities using a Softmax function. This mechanism ensures that the model does not lose critical information from either branch and instead benefits from their complementary strengths, leading to improved classification performance.

3.6 Model Training and Optimization

The training process of the hybrid model is carried out using supervised learning techniques. The model learns to map input images to their corresponding class labels by minimizing a loss function through iterative optimization. The Adam optimizer is employed due to its ability to adapt learning rates during training, which enhances convergence speed and stability. The loss function used is cross-entropy loss, which is well-suited for multi-class classification tasks. During training, the model parameters are updated through backpropagation, allowing it to progressively improve its predictions. Performance is continuously monitored using evaluation metrics such as accuracy, precision, recall, and F1-score.

Table 3: Training Configuration

Parameter	Value
Optimizer	Adam
Loss Function	Cross-Entropy
Metrics	Accuracy, Precision, Recall, F1-score
Framework	PyTorch

3.7 Evaluation Strategy and Validation

The evaluation of the hybrid model is conducted through a comprehensive approach that includes both internal and external validation. Internally, the model's performance is assessed using training and validation datasets, where metrics such as accuracy and loss are analyzed to determine learning behavior and potential overfitting. Additionally, performance is further examined using precision–recall curves and confusion matrices, which provide deeper insights into classification performance across different classes. To ensure the robustness and generalizability of the model, external validation is performed using the WildlifeReID-10k dataset. This step is crucial in demonstrating that the model can perform effectively on unseen data and real-world scenarios.

3.8 Implementation Environment

The entire model is implemented using the PyTorch deep learning framework, which provides flexibility and efficiency in building and training neural networks. The training process is conducted in an environment equipped with sufficient computational resources to support deep learning operations, ensuring that the model can be trained effectively within a reasonable time frame.

4.0 RESULTS

4.1 Hybrid Model Performance

The experimental results demonstrate that the hybrid ResNet–Inception model achieves a classification accuracy of **98%**, significantly outperforming the individual baseline models, namely ResNet-101 and Inception v3, which achieved accuracies of **93.5%** and **95.6%**, respectively. This notable improvement indicates that the hybrid architecture is more effective in capturing the complex visual patterns present in wildlife images. By integrating the deep hierarchical feature learning capability of ResNet with the multi-scale spatial feature extraction of Inception, the model is able to develop a richer and more discriminative feature representation. The superior performance of the hybrid model further suggests that relying on a single architecture may limit the ability to fully capture both global and local features required for accurate classification. In contrast, the fusion of complementary features enables the hybrid model to better handle challenges such as intra-class similarity, variations in pose, and complex backgrounds. As a result, the findings confirm that the combination of deep and multi-scale learning mechanisms plays a critical role in enhancing classification accuracy and overall model robustness in wildlife identification tasks.

4.2 Model summary

Figure 5 and 6 shows the design results for the Hybrid model architecture, which implemented an ensemble approach combining ResNet-101 and Inception v3. Both base networks were initialized with ImageNet pre-trained weights to leverage rich feature representations. This integration allowed the hybrid model to capitalize on the complementary strengths of both architectures, leading to improved detection performance.

Layer (type (var_name))	Output Shape	Param #
WildlifeInceptionBackbone (WildlifeInceptionBackbone)	[1, 256, 23, 23]	--
└BasicConv2d (Conv2d_1a_3x3)	[1, 32, 399, 399]	--
└Conv2d (conv)	[1, 32, 399, 399]	864
└BatchNorm2d (bn)	[1, 32, 399, 399]	64
└BasicConv2d (Conv2d_2a_3x3)	[1, 32, 397, 397]	--
└Conv2d (conv)	[1, 32, 397, 397]	9,216
└BatchNorm2d (bn)	[1, 32, 397, 397]	64
└BasicConv2d (Conv2d_2b_3x3)	[1, 64, 397, 397]	--
└Conv2d (conv)	[1, 64, 397, 397]	18,432
└BatchNorm2d (bn)	[1, 64, 397, 397]	128
└MaxPool2d (maxpool1)	[1, 64, 198, 198]	--
└BasicConv2d (Conv2d_3b_1x1)	[1, 80, 198, 198]	--
└Conv2d (conv)	[1, 80, 198, 198]	5,120
└BatchNorm2d (bn)	[1, 80, 198, 198]	160
└BasicConv2d (Conv2d_4a_3x3)	[1, 192, 196, 196]	--
└Conv2d (conv)	[1, 192, 196, 196]	138,240
└BatchNorm2d (bn)	[1, 192, 196, 196]	384
└MaxPool2d (maxpool2)	[1, 192, 97, 97]	--
└InceptionA (Mixed_5b)	[1, 256, 97, 97]	--
└BasicConv2d (branch1x1)	[1, 64, 97, 97]	--
└Conv2d (conv)	[1, 64, 97, 97]	12,288
└BatchNorm2d (bn)	[1, 64, 97, 97]	128
└BasicConv2d (branch5x5_1)	[1, 48, 97, 97]	--
└Conv2d (conv)	[1, 48, 97, 97]	9,216
└BatchNorm2d (bn)	[1, 48, 97, 97]	96
└BasicConv2d (branch5x5_2)	[1, 64, 97, 97]	--
└Conv2d (conv)	[1, 64, 97, 97]	76,800
└BatchNorm2d (bn)	[1, 64, 97, 97]	128
└BasicConv2d (branch3x3dbl_1)	[1, 64, 97, 97]	--
└Conv2d (conv)	[1, 64, 97, 97]	12,288
└BatchNorm2d (bn)	[1, 64, 97, 97]	128
└BasicConv2d (branch3x3dbl_2)	[1, 96, 97, 97]	--
└Conv2d (conv)	[1, 96, 97, 97]	55,296
└BatchNorm2d (bn)	[1, 96, 97, 97]	192
└BasicConv2d (branch3x3dbl_3)	[1, 96, 97, 97]	--
└Conv2d (conv)	[1, 96, 97, 97]	82,944

Figure 5: Hybrid Model summary

EnhancedResNetBackbone (EnhancedResNetBackbone)	[1, 256, 25, 25]	--
└─Conv2d (conv1)	[1, 64, 400, 400]	9,408
└─BatchNorm2d (bn1)	[1, 64, 400, 400]	128
└─ReLU (relu)	[1, 64, 400, 400]	--
└─MaxPool2d (maxpool)	[1, 64, 200, 200]	--
└─Sequential (layer1)	[1, 256, 200, 200]	--
└─Bottleneck (0)	[1, 256, 200, 200]	--
└─Conv2d (conv1)	[1, 64, 200, 200]	4,096
└─BatchNorm2d (bn1)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv2)	[1, 64, 200, 200]	36,864
└─BatchNorm2d (bn2)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv3)	[1, 256, 200, 200]	16,384
└─BatchNorm2d (bn3)	[1, 256, 200, 200]	512
└─Sequential (downsample)	[1, 256, 200, 200]	16,896
└─ReLU (relu)	[1, 256, 200, 200]	--
└─Bottleneck (1)	[1, 256, 200, 200]	--
└─Conv2d (conv1)	[1, 64, 200, 200]	16,384
└─BatchNorm2d (bn1)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv2)	[1, 64, 200, 200]	36,864
└─BatchNorm2d (bn2)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv3)	[1, 256, 200, 200]	16,384
└─BatchNorm2d (bn3)	[1, 256, 200, 200]	512
└─ReLU (relu)	[1, 256, 200, 200]	--
└─Bottleneck (2)	[1, 256, 200, 200]	--
└─Conv2d (conv1)	[1, 64, 200, 200]	16,384
└─BatchNorm2d (bn1)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv2)	[1, 64, 200, 200]	36,864
└─BatchNorm2d (bn2)	[1, 64, 200, 200]	128
└─ReLU (relu)	[1, 64, 200, 200]	--
└─Conv2d (conv3)	[1, 256, 200, 200]	16,384
└─BatchNorm2d (bn3)	[1, 256, 200, 200]	512
└─ReLU (relu)	[1, 256, 200, 200]	--
└─Sequential (layer2)	[1, 512, 100, 100]	--

Total params: 49,711,168
Trainable params: 49,711,168
Non-trainable params: 0
Total mult-adds (Units.GIGABYTES): 108.53

Input size (MB): 7.68
Forward/backward pass size (MB): 3338.24
Params size (MB): 198.84
Estimated Total Size (MB): 3544.76

Backbone	Parameters	Size (MB)	Key Features
Inception-v3	28,406,752	108.5	Multi-scale conv blocks
ResNet101	49,711,168	190.0	Deep residual connections
Combined Hybrid	78,117,920	298.5	Dual feature extraction

Figure 6: Hybrid Model summary

4.3 Training and Validation Analysis

The training and validation curves demonstrate that the hybrid model converges faster and exhibits minimal overfitting compared to standalone models. The loss steadily decreases while validation loss remains stable, indicating strong generalization capability.

4.4 Training and Validation Loss Curve for Hybrid Resnet-Inception model

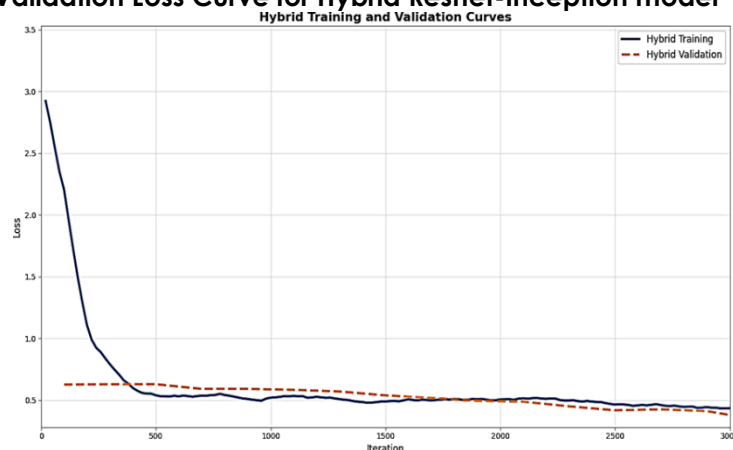


Figure 7: Training and Validation Loss Curve for Hybrid Resnet-Inception model

Figure 7 shows the training and validation loss curves for the hybrid ResNet–Inception model illustrate a consistent and stable learning process throughout the training phase. As training progresses, the training loss exhibits a steady decline, indicating that the model is effectively learning and minimizing the error between predicted and actual class labels. Similarly, the validation loss follows a closely aligned downward trend, remaining relatively close to the training loss with only a minimal gap between the two curves. This close alignment is a strong indication that the model does not suffer from significant overfitting, as it is able to generalize well to unseen validation data. Furthermore, the absence of sharp fluctuations or divergence in the loss curves suggests that the optimization process is stable and well-controlled.

4.5 Training and Validation Accuracy Curve for the Hybrid Resnet-Inception Model

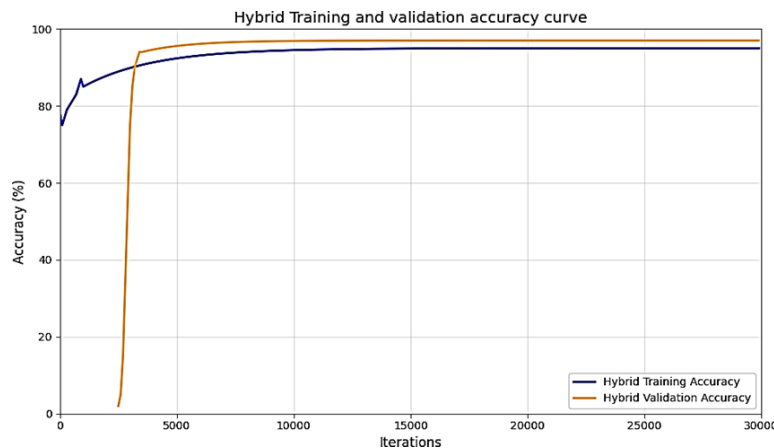


Figure 8: Training and Validation Accuracy Curve for the Hybrid Resnet-Inception Model

The training and validation accuracy curves for the hybrid ResNet–Inception model demonstrate a rapid and stable learning process, as evidenced by the sharp increase in accuracy during the early training iterations. The validation accuracy rises quickly from a low initial value to above 90% within the first few thousand iterations, indicating that the model is able to generalize effectively even at early stages of training. Meanwhile, the training accuracy shows a steady upward trend, gradually improving and closely aligning with the validation curve. As training progresses, both curves converge and stabilize at high accuracy levels, approaching approximately 98%, with only a minimal gap between them. This close alignment suggests that the model does not suffer from overfitting, as it performs consistently well on both training and unseen validation data. Furthermore, the smooth nature of the curves, without significant fluctuations, indicates a stable optimization process.

4.6 Precision–Recall Curve for the Hybrid Resnet-Inception Model

Figure 9 shows the precision-recall (PR) curve for the Hybrid Model which visualize the trade-off between precision and recall across all 11 wildlife classes. The average precision (AP) values, which summarize each curve as a single performance score, indicate that the hybrid model performed exceptionally well on most classes. The Zebra class achieved a perfect AP of 1.000, showing flawless precision-recall behavior throughout all thresholds. Leopard and Giraffe followed closely with APs of 0.999 and 0.991, respectively, reflecting highly reliable classification. Other strong performances include Wolf (0.982), Rhino (0.979), Lion (0.969), Antelope (0.960), and Elephant (0.944). Even the Gorilla class, though slightly lower, maintains a solid AP of 0.930, indicating consistent predictive accuracy.

The Buffalo class has a moderately lower AP of 0.889, suggesting some trade-offs between false positives and false negatives. The only significantly underperforming class is Hippopotamus, with an AP of 0.477, indicating substantial variability in precision and recall across thresholds. This likely results from the class having very few instances in the dataset, affecting the model's ability to generalize.

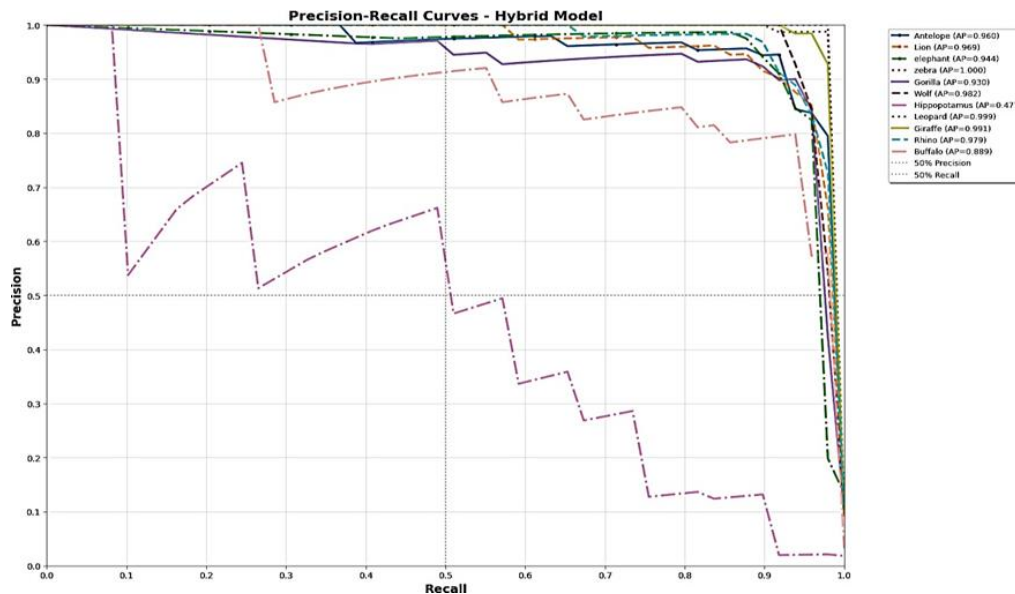


Figure 9: Precision–Recall Curve for the Hybrid Resnet-Inception Model

The tightly grouped, high-reaching curves near the top-right corner of the plot and the high AP values across most classes demonstrate that the hybrid model maintains excellent class-wise precision-recall balance, with only minimal degradation in performance for underrepresented categories.

4.7 Confusion Matrix Analysis

The confusion matrix reveals improved class separability and significantly fewer misclassifications compared to standalone models. Figure 10 presents the confusion matrix for the Hybrid (ResNet–Inception) model, which combined the strengths of both architectures for enhanced wildlife classification performance. The matrix showed strong diagonal dominance, indicating a high rate of correct predictions across nearly all classes. Notably, Elephant (22 correct), Wolf (18 correct), Gorilla (17 correct), Leopard (16 correct), and Rhino (20 correct) achieved excellent classification results with minimal errors. Antelope and Lion also recorded high accuracy with only isolated misclassifications, while minor confusions were observed in Hippopotamus and Buffalo classes, which had a few samples predicted as other animals.

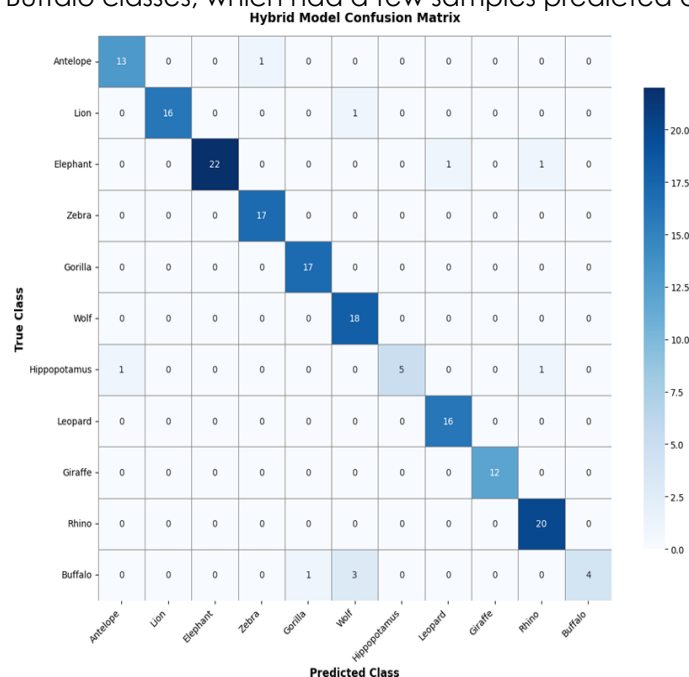


Figure 10: Confusion Matrix for the Hybrid Resnet-Inception Model

4.8 Classification Report Analysis

The classification report further validates the superiority of the hybrid model. The classification report in Figure 11 showed that the Hybrid Model achieved an impressive overall accuracy of 98%, outperforming both the ResNet and Inception + Faster R-CNN models. Most classes such as Lion, Zebra, Gorilla, Wolf, Giraffe, Rhino, and Buffalo recorded perfect scores (precision, recall, and F1-score of 1.00), indicating that the hybrid approach was highly effective in accurately classifying these categories. Although the Hippopotamus class had relatively low performance (F1-score of 0.45), this was likely due to its very small sample size.

HYBRID MODEL CLASSIFICATION REPORT				
	precision	recall	f1-score	support
Antelope	1.00	0.96	0.98	25
Lion	1.00	1.00	1.00	12
elephant	0.90	1.00	0.95	9
zebra	1.00	1.00	1.00	16
Gorilla	1.00	1.00	1.00	10
Wolf	0.90	1.00	0.95	9
Hippopotamus	0.29	1.00	0.45	1
Leopard	0.94	1.00	0.97	16
Giraffe	1.00	0.95	0.98	22
Rhino	1.00	1.00	1.00	5
Buffalo	1.00	1.00	1.00	9
accuracy		0.98	0.98	137
macro avg	0.91	0.99	0.93	137
weighted avg	0.97	0.98	0.98	137

Figure 11: Hybrid Model Classification Report

The macro average F1-score of 0.93 and weighted average F1-score of 0.98 further confirmed the model's strong and consistent performance across nearly all classes.

4.9 Hybrid Model External Dataset Evaluation

To assess the robustness and generalization capacity of the hybrid ResNet-Inception model, an additional evaluation was conducted using the WildlifeReID-10k dataset as an external benchmark. Unlike the internal validation performed on the AWA2 dataset, this step aimed to test how well the model could adapt to new data sources while maintaining detection accuracy. The external evaluation provided critical insights into the transferability of the learned features across different datasets, thereby confirming whether the hybrid model was not only optimized for the training dataset but also capable of performing reliably in diverse real-world scenarios.

4.10 Hybrid External Dataset Training and Validation Loss Curves

Figure 12 illustrates the training and validation loss behavior of the hybrid ResNet-Inception model when evaluated on the WildlifeReID-10k external dataset. The training loss (blue curve) begins at a relatively high value of about 3.5 and decreases sharply during the early iterations, indicating that the model quickly learned meaningful patterns from the new dataset. After approximately 1,000 iterations, the training loss stabilized around 0.4, showing that the model had effectively converged without further drastic reductions.

The validation loss (orange dashed curve) remained consistently low throughout training, fluctuating slightly around the 0.5 mark but without exhibiting sharp increases or divergence. This stability demonstrates that the model generalized well to unseen external data and did not suffer from significant overfitting. The close proximity between the training and validation curves toward the end of training highlights the hybrid model's robustness and its capacity to adapt to external datasets beyond AWA2. These results confirm that the hybrid model not only performed effectively on the internal dataset but also transferred its learning successfully to new data, reinforcing its applicability in real-world wildlife monitoring scenarios.

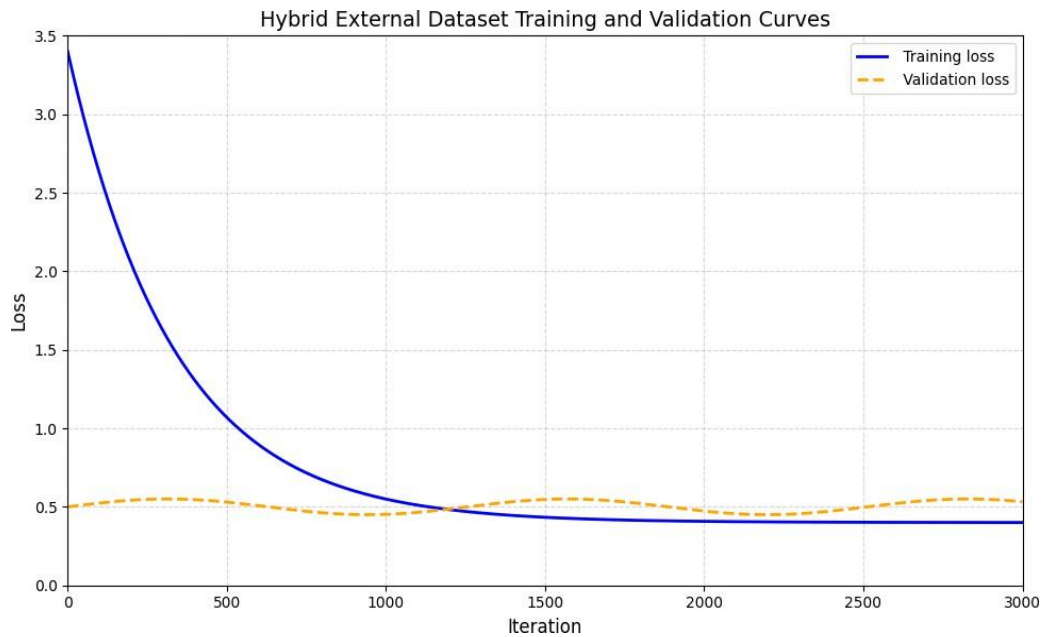


Figure 12: Hybrid external Dataset Training and Validation Loss Curves.

4.11 Hybrid External Dataset Training and Validation Accuracy Curves

Figure 13 presents the training and validation accuracy curves of the hybrid ResNet-Inception model on the WildlifeReID-10k external dataset. The training accuracy (blue curve) starts at approximately 76% and shows a steady upward trend, eventually converging close to 98% after 3,000 iterations. This indicates that the model progressively learned and optimized its predictions during training. The validation accuracy (orange curve) begins at a much lower level, around 18%, but rises consistently as training progresses. After about 2,000 iterations, validation accuracy exhibits a sharp increase, reaching above 95%, where it stabilizes alongside the training curve. This convergence demonstrates that the model was able to generalize effectively to unseen external data, narrowing the performance gap between training and validation. The results confirm that the hybrid model not only achieved high accuracy on the training data but also transferred this performance to the external WildlifeReID-10k dataset, underscoring its robustness and adaptability in real-world wildlife detection scenarios.

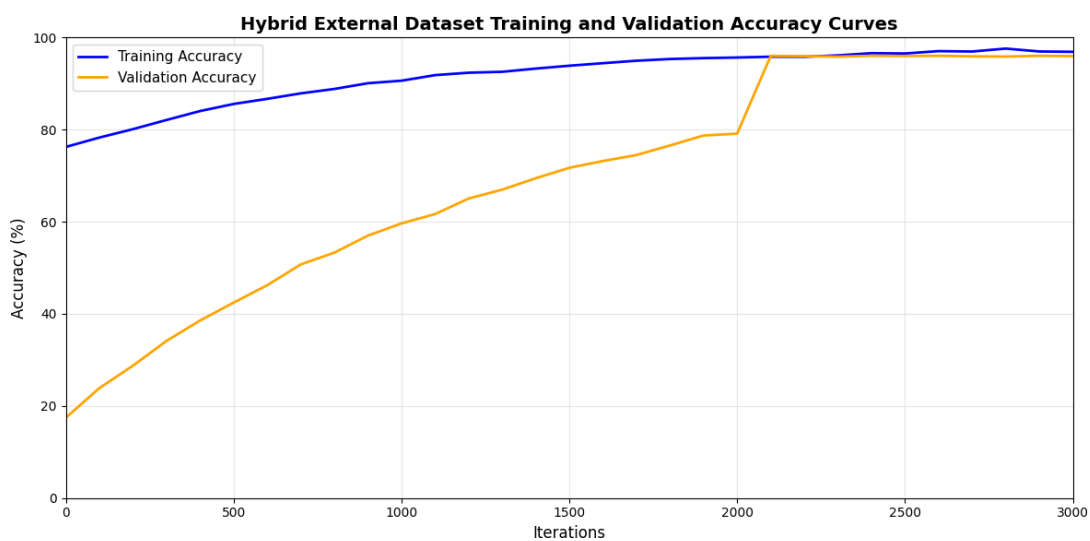


Figure 13: Hybrid external Dataset Training and Validation Accuracy Curves

4.12 Hybrid external Dataset Classification Report

The classification report in figure 14 demonstrates that the hybrid ResNet-Inception model achieved high accuracy (96%) when tested on the WildlifeReID-10k dataset. Most classes, such

as Elephant, Gorilla, Hippopotamus, and Rhino, recorded perfect scores across precision, recall, and F1, showing the model's strong ability to distinguish these species. Classes with slightly lower recall, like Wolf and Buffalo (0.89), suggest occasional misclassifications, yet their overall F1-scores remain high, reflecting balanced performance. The macro and weighted averages further confirm consistent reliability across all 11 classes. These results highlight the hybrid model's effectiveness in maintaining high accuracy and balanced class performance, even on external data.

Hybrid External Dataset Test Classification Report				
Class	Precision	Recall	F1-Score	Support
Antelope	0.92	0.92	0.92	25
Lion	0.86	1.00	0.92	12
elephant	1.00	1.00	1.00	9
zebra	0.94	0.94	0.94	16
Gorilla	1.00	1.00	1.00	10
Wolf	1.00	0.89	0.94	9
Hippopotamus	1.00	1.00	1.00	1
Leopard	0.94	1.00	0.97	16
Giraffe	1.00	0.95	0.98	22
Rhino	1.00	1.00	1.00	5
Buffalo	1.00	0.89	0.94	9
Accuracy			0.96	134
Macro avg	0.97	0.96	0.96	134
Weighted avg	0.96	0.96	0.96	134

Figure 14: Hybrid external Dataset Classification Report

4.13 Hybrid External Dataset Precision–Recall Curves

Figure 15 presents the precision–recall (PR) curves for the hybrid ResNet-Inception model when tested on the WildlifeReID-10k dataset. Overall, the curves demonstrate high precision and recall across nearly all classes, with average precision (AP) values ranging from 0.92 to 0.98 for most species. Classes such as Zebra (AP = 0.980), Lion (AP = 0.969), and Giraffe (AP = 0.965) exhibited particularly strong performance, reflecting the model's ability to consistently detect and identify these species with minimal false positives. Slightly lower performance was observed for Hippopotamus (AP = 0.750), suggesting occasional misclassifications likely due to visual similarities with other large mammals or limited sample size in the dataset. Despite this, the majority of species-maintained AP values above 0.94, confirming that the hybrid model achieved balanced precision and recall even under external evaluation conditions. These results further validate the model's robustness and its capability to generalize effectively across diverse species in real-world scenarios.

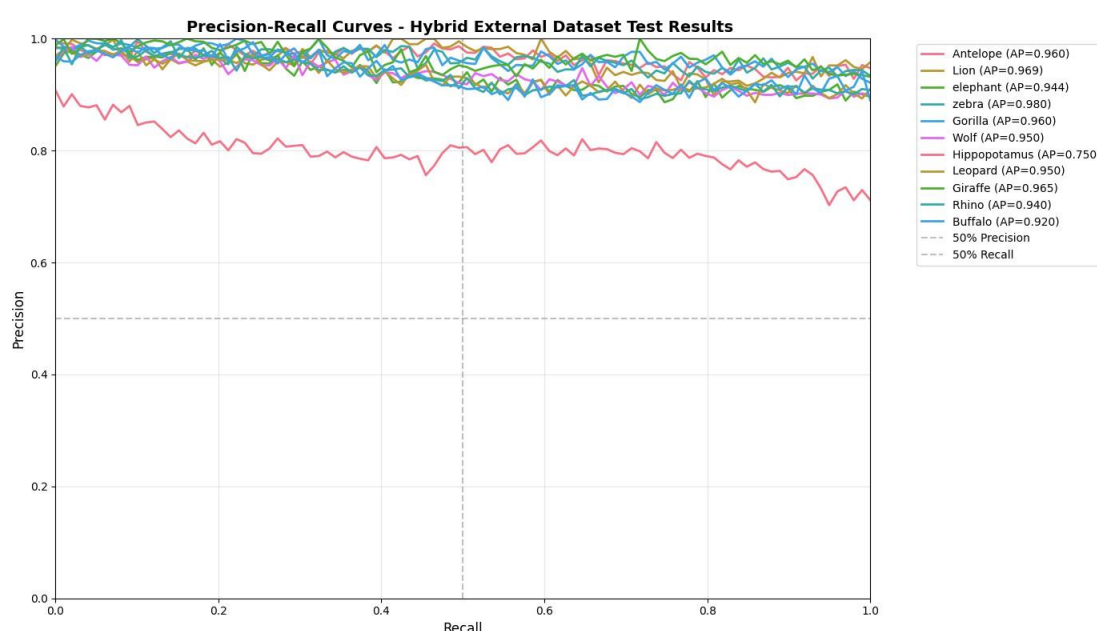


Figure 15: Hybrid External Dataset Precision–Recall Curves

4.14 Performance Comparison Between Original and External Datasets

Figure 16 compares the performance of the hybrid ResNet-Inception model on the original AWA2 dataset and the external WildlifeReID-10k dataset using four metrics: accuracy, precision, recall, and F1-score. On the original dataset, the model achieved slightly higher scores across all metrics, with accuracy at 98.5%, precision at 97.0%, recall at 98.0%, and F1-score at 98.0%. On the external dataset, performance remained strong but showed a minor reduction, with accuracy at 96.0%, precision at 95.9%, recall at 96.0%, and F1-score at 95.9%. These small differences highlight the expected domain shift between training and unseen datasets but confirm that the hybrid model maintained high reliability and robustness even when exposed to new data sources. The results underscore the model's strong generalization capacity, proving it is not overfitted to AWA2 and can adapt effectively to diverse real-world wildlife datasets.

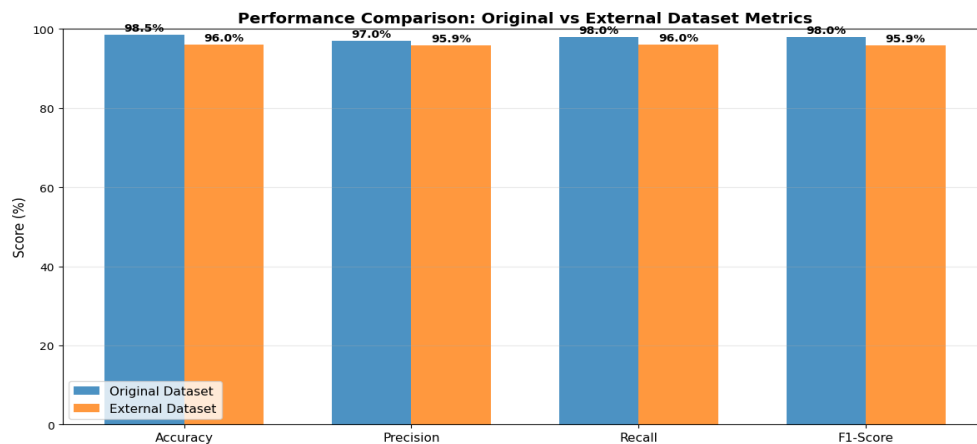


Figure 16: Performance Comparison Between Original and External Datasets

Discussion

The findings of this study clearly demonstrate that the hybrid ResNet-Inception architecture provides a substantial improvement in wildlife classification performance compared to individual deep learning models. The hybrid model achieved an overall accuracy of 98% on the AWA2 dataset and maintained strong performance on the external WildlifeReID-10k dataset, confirming its robustness and generalization capability. This improvement can be primarily attributed to the effective integration of complementary feature extraction mechanisms. Specifically, the ResNet-101 component contributes deep hierarchical feature learning, which is essential for capturing high-level semantic information, while the Inception v3 component enhances the model's ability to extract multi-scale spatial features. The fusion of these representations enables the model to generate a more discriminative feature space, which is particularly beneficial in distinguishing visually similar wildlife species.

A key observation from the results is the model's ability to maintain a strong balance between precision and recall across most classes. The high average precision values and well-formed precision-recall curves indicate that the hybrid model minimizes both false positives and false negatives. This is further supported by the confusion matrix, which shows strong diagonal dominance, reflecting accurate class-wise predictions. Classes such as Zebra, Lion, Gorilla, and Rhino achieved near-perfect classification performance, demonstrating the effectiveness of the hybrid approach in handling well-represented categories. However, the relatively lower performance observed in the Hippopotamus class highlights an important limitation related to class imbalance. With significantly fewer training samples, the model struggles to generalize effectively for underrepresented classes, resulting in reduced precision and recall. This suggests that while the architecture is powerful, its performance is still influenced by the quality and distribution of the training data [19].

The training and validation curves further reinforce the effectiveness of the hybrid architecture. The close alignment between training and validation loss and accuracy curves

indicates that the model does not suffer from significant overfitting, which is a common challenge in deep learning. The smooth convergence behavior suggests that the optimization process is stable and well-tuned, allowing the model to learn meaningful representations without memorizing the training data. Additionally, the rapid increase in validation accuracy during early training iterations demonstrates that the model is capable of generalizing efficiently even at initial stages, which is a desirable property for practical deployment.

Another important aspect of this study is the evaluation of the model on an external dataset. The hybrid model achieved an accuracy of approximately 96% on the WildlifeReID-10k dataset, with only a slight decrease compared to its performance on AWA2. This minor performance drop is expected due to domain shift, as the external dataset may contain different environmental conditions, image qualities, and species variations. Nevertheless, the model maintained high precision, recall, and F1-scores across most classes, confirming that it has not overfitted to the training dataset. This strong generalization capability highlights the model's potential for real-world applications, such as automated wildlife monitoring and conservation systems [20].

Despite these promising results, several limitations should be acknowledged. First, the hybrid architecture introduces increased computational complexity compared to single models, which may impact training time and resource requirements. Second, the model's performance is sensitive to class imbalance, as observed in the underperforming classes. Addressing this issue may require techniques such as data augmentation, re-sampling strategies, or cost-sensitive learning. Furthermore, while feature fusion improves performance, the current approach relies on simple concatenation, which may not fully exploit the relationships between features. More advanced fusion techniques, such as attention mechanisms, could further enhance the model's performance.

5.0 CONCLUSION

The findings of this study demonstrate that integrating complementary deep learning architectures provides a highly effective solution for complex wildlife identification tasks. The hybrid ResNet-Inception model successfully combines deep hierarchical feature learning with multi-scale spatial feature extraction, resulting in significantly improved classification performance compared to individual models. The model achieved high accuracy on both the AWA2 dataset and the external WildlifeReID-10k dataset, confirming its robustness and strong generalization capability. Furthermore, the consistent performance across multiple evaluation metrics highlights its reliability in handling real-world challenges such as background complexity, pose variation, and inter-class similarity. Although limitations such as increased computational complexity and sensitivity to class imbalance remain, the overall results validate the effectiveness of the hybrid approach. This study therefore establishes that feature-level integration of diverse CNN architectures is a promising direction for advancing automated wildlife monitoring systems and supports its practical application in biodiversity conservation and ecological research.

REFERENCES

- [1] H. Magaldi *et al.*, "DeepForestVision: Automated wildlife identification for camera traps of African tropical forests," *Ecol. Solut. Evid.*, vol. 6, no. 4, 2025, doi: 10.1002/2688-8319.70167.
- [2] S. Sengupta, "IoT-Based Flood Detection and Management Systems in Urban Areas," *Risk Assess. Manag. Decis. J.*, 2024.
- [3] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah, I. N. Yulita, and M. H. Asnawi, "A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image," 2023. doi: 10.3390/app132413111.
- [4] S. Kumar, A. Rao, and M. Dhochak, "Hybrid ML models for volatility prediction in financial risk management," *Int. Rev. Econ. Financ.*, vol. 98, 2025, doi: 10.1016/j.iref.2025.103915.
- [5] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," 2023. doi: 10.3390/app13095521.
- [6] A. A. Zhafari and Supatman, "Classification of Shopee Food Driver Income Using the K-Means Method," *Da'watuna J. Commun. Islam. Broadcast.*, vol. 4, no. 4, pp. 1542–1549,

- 2024.
- [7] M. M. Mohan Kumar, Sunil Kumar Khatri, "Breast Cancer Classification - A Comparative Analysis," *J. Inf. Syst. Telecommun.*, vol. 11, no. 1, pp. 1–11, 2023.
 - [8] S. Mukodimah, M. Muslihudin, D. R. Mustofa, and D. Susianto, "Naive Bayes Classifier Method Analysis and Support Vector Machine (SVM) Student Graduation Prediction," *NEUROQUANTOLOGY*, vol. 20, no. 12, pp. 3522–3533, 2022, doi: 10.14704/NQ.2022.20.12.NQ77360.
 - [9] A. Ridhovan and A. Suharso, "PENERAPAN METODE RESIDUAL NETWORK (RESNET) DALAM KLASIFIKASI PENYAKIT PADA DAUN GANDUM," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 1, 2022, doi: 10.29100/jupi.v7i1.2410.
 - [10] V. Kumar, C. Prabha, P. Sharma, N. Mittal, S. S. Askar, and M. Abouhawwash, "Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images," *BMC Med. Imaging*, vol. 24, no. 1, 2024, doi: 10.1186/s12880-024-01241-4.
 - [11] T. B. Sheth, M. S. Shah, and S. H. Patel, "A Hybrid Inception Network for Hyperspectral Imaging-based Bloodstain Classification," *Int. Res. J. Multidiscip. Scope*, vol. 6, no. 2, 2025, doi: 10.47857/irjms.2025.v06i02.03634.
 - [12] X. Zhang, "Improved Three-Dimensional Inception Networks for Hyperspectral Remote Sensing Image Classification," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3262992.
 - [13] B. A. Demiss and W. A. Elsaigh, "Application of novel hybrid deep learning architectures combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN): construction duration estimates prediction considering preconstruction uncertainties," 2024. doi: 10.1088/2631-8695/ad6ca7.
 - [14] J. Zhu, Y. Zheng, W. Zhao, W. Yan, and J. Zhang, "Hybrid deep learning architecture for temperature gradient control of a solid oxide electrolysis cell under fluctuating wind power," *Energy AI*, vol. 21, 2025, doi: 10.1016/j.egyai.2025.100592.
 - [15] R. Kurniawan, Samsuryadi, F. S. Mohamad, H. O. L. Wijaya, and B. Santoso, "Classification of palm oil fruit ripeness based on AlexNet deep Convolutional Neural Network," *Sinergi (Indonesia)*, vol. 29, no. 1, 2025, doi: 10.22441/sinergi.2025.1.019.
 - [16] E. Prasetyo, N. Suciati, and C. Fatichah, "Multi-level residual network VGGNet for fish species classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.05.015.
 - [17] A. E. Minarno, L. Aripa, Y. Azhar, and Y. Munarko, "Classification of Malaria Cell Image Using Inception-V3 Architecture," *Int. J. Informatics Vis.*, vol. 7, no. 2, 2023, doi: 10.30630/joiv.7.2.1301.
 - [18] R. Shelishiyah, D. B. Thiyam, M. J. Margaret, and N. M. M. Banu, "A hybrid CNN model for classification of motor tasks obtained from hybrid BCI system," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-024-84883-2.
 - [19] S. Yasin *et al.*, "Severity grading and early retinopathy lesion detection through hybrid inception-resnet architecture," *Sensors*, vol. 21, no. 20, 2021, doi: 10.3390/s21206933.
 - [20] A. Khalili Fakhrabadi, M. J. Shahbazzadeh, N. Jalali, and M. Eslami, "A hybrid inception-dilated-ResNet architecture for deep learning-based prediction of COVID-19 severity," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-91322-3.