AI-based Wildlife Species Identification

Classify Animals in Camera Trap Images

Student Name: Khadija Saeed   Registration No.: 2023-CS-74
Supervisor Name: Sir Waseem
Department of Computer Science, University of Engineering & Technology, Lahore

Abstract

This work presents a wildlife species detection system based on the Vision Transformer (ViT) architecture. The model was fine-tuned using the Animals-10 dataset, containing over 28,000 images across 10 animal categories. The proposed system achieved 98.04% accuracy, improving overall accuracy by 6.4% compared to previous CNN-based approaches and showing a 22% improvement in detecting hard-to-identify species under occlusion and low light.
The model was implemented using PyTorch and Hugging Face Transformers. Performance was monitored using training and validation loss and accuracy. Confusion matrices and interpretability tools were used to evaluate decision-making behavior. The system can support conservation efforts and send safety alerts in areas where wild animals may enter human environments.

Introduction

Background

In many parts of the world, animals and humans live close to each other, especially near forests and wildlife reserves. This often leads to serious problems, like attacks from wild animals (e.g., tigers, elephants), damage to crops by wild boars, and the spread of diseases from animals to humans. At the same time, many animal species are at risk of extinction due to habitat destruction and illegal hunting.

Motivation

Fast and accurate identification of animals can help solve both wildlife and human safety issues. However, current wildlife monitoring methods mostly depend on experts reviewing images manually, which is slow and costly. Even automated systems using CNNs often struggle in real-life conditions, like poor lighting or when animals are partially hidden.

Research Objectives and Questions

This study aims to develop a smart system that can:

- Identify animals accurately, even in difficult conditions (e.g., night or camouflage),
- Work fast enough to send real-time alerts to nearby villages.

Key research questions include:

- Can we improve animal identification accuracy using modern deep learning models?
- Is the system efficient enough to run on low-power devices in remote areas?
- Can it be useful for both conservation and human safety?

Significance of the Study

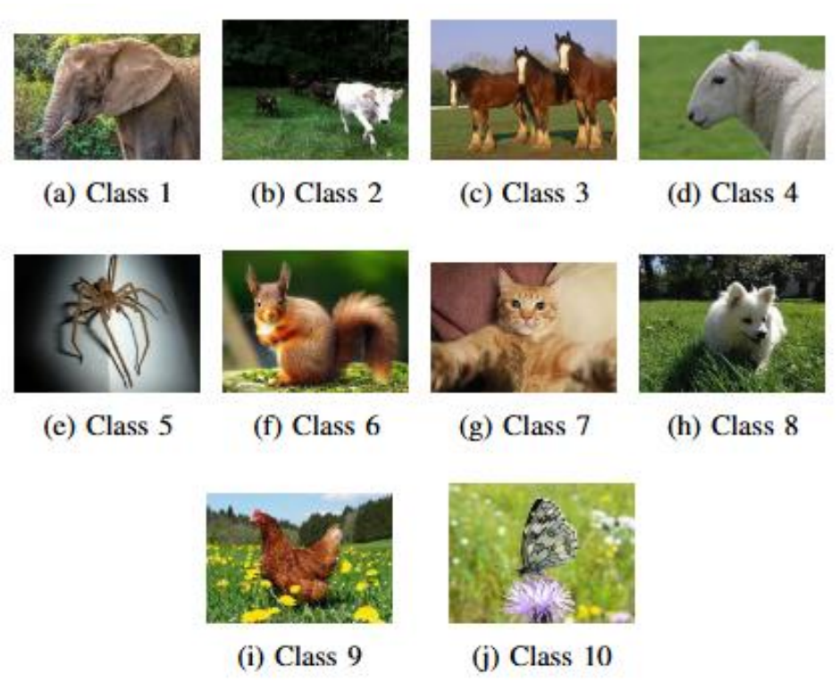
Our system uses a Vision Transformer model fine-tuned on the Animals-10 dataset. It achieves a **high accuracy of 98.04%**, even in poor lighting or when animals are partially hidden. This makes it more reliable than traditional CNN-based systems. The model is also lightweight enough to run on edge devices in remote areas.
It offers two major benefits:
1. **Helping conservationists** monitor endangered species more effectively.
2. **Protecting communities** by sending timely alerts when dangerous animals are detected nearby.

Related Work

Year	Paper	Model	Dataset	Limitation	Result
2018	Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning	Deep CNN	Camera traps	Imbalanced data, limited multi-species handling, noisy labels, and reliance on classification for counting.	Accuracy: 93.8%
2022	Identification and Classification of Wild Animals from Video Sequences Using Hybrid Deep Residual Convolutional Neural Network	Hybrid CNN	Videos	Performance may degrade with highly occluded or low-resolution animal images in complex forest backgrounds.	Accuracy: 97.35%
2023	Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images	Deep CNN	Ecological Camera Traps	Imbalanced dataset, which risks model bias and poor generalization to images from new locations or conditions.	Accuracy: 89.2%
2023	Multimodal Foundation Models for Zero-shot Animal Species Recognition in Camera Trap Images	VLM	Camera Traps dataset	High computational cost, limits its efficiency compared to supervised models.	Accuracy: 85.6%
2023	Deep Learning Based Bird Species Identification and Classification Using Images	Skip-CNN	Birds	Relatively small number of images per species.	Accuracy: 92.0%
2024	Transformer-Based Wildlife Species Classification Using High-Frequency Features	Transformer	Eco. Parks	Risk of model bias.	Accuracy: 92.7%
2024	Large-Scale Multispecies Animal Re-identification Using Community-Curated Data	EffNetV2	Multiple Datasets	The uneven dataset lowers accuracy for rare species.	Accuracy: 93.4%
2024	Dual-Channel CNN for Wildlife Monitoring in Biodiversity Conservation	Dual-CNN	Wildlife	High computational cost degrades the model's efficiency.	Accuracy: 95.2%
2024	Adaptive High-Frequency Transformer for Diverse Wildlife Re-Identification	HF-Transformer	Multi-dataset	Performance may get affected due to imbalance dataset.	Accuracy: 94.1%
2024	Towards Context-Rich Automated Biodiversity Assessments	YOLOv10-X	Camera traps	Rely on clear bounding box labels for species identification	Accuracy: 90.3% F1 score: 0.96

Methodology

1. Dataset and Preprocessing


(a) Class 1 (b) Class 2 (c) Class 3 (d) Class 4
(e) Class 5 (f) Class 6 (g) Class 7 (h) Class 8
(i) Class 9 (j) Class 10

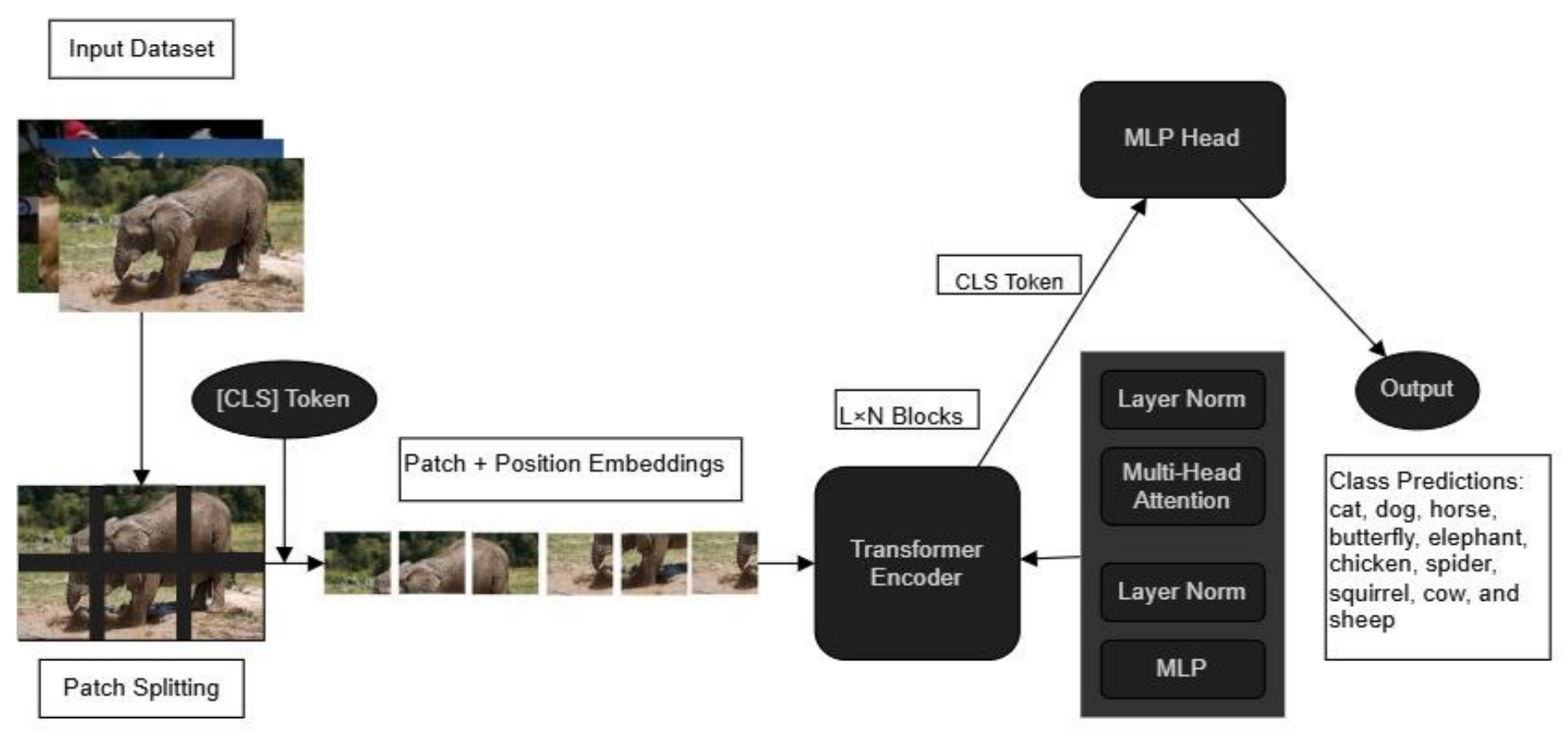
The Animals-10 dataset comprises approximately28,000 medium-quality image. The dataset was split into three parts: **70% for training, 15% for validation, and 15% for testing**. The image pixel values were scaled to the range [0, 1] by dividing by 255:

Rescaled Image = Original Pixel Value/255

Data augmentation techniques, including random horizontal flipping, rotation, and color jittering, were applied to increase data diversity and improve generalization.

2. Model Architecture

We employed a Vision Transformer (ViT-B/16) architecture, pre-trained on the ImageNet-21k dataset. Unlike convolutional neural networks, ViT splits the input image into patches, embeds each patch linearly, and processes them using self-attention layers. This allows the model to capture long-range dependencies and contextual information effectively.



2. Training Procedure

The model was trained using the Adam optimizer with a **learning rate of 2×10−5** and a **batch size of 32**. Cross-entropy loss was used as the objective function. Early stopping and learning rate scheduling were employed to prevent overfitting and optimize convergence. Training was conducted for 12epochs using PyTorch on a single NVIDIA GPU (TeslaT4).

Methodology (Continue)

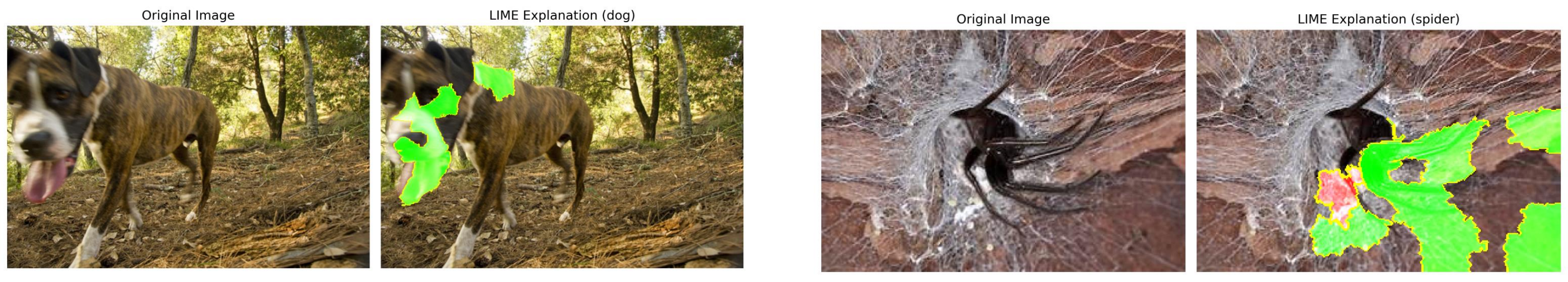
Evaluation Matrices

Model performance was evaluated using accuracy, precision, recall, and F1-score. A confusion matrix was also plotted to analyze class-wise prediction behavior. These metrics were computed on the held-out test set to ensure unbiased evaluation. The proposed method aims to achieve high classification performance across all species, with emphasis on minimizing misclassification between visually similar classes such as dog and wolf, or sheep and cow.

- **Accuracy:**
$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100$$
- **Precision:**
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$
- **Recall:**
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$
- **F1-Score:**
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

Model Explanation (LIME)

To understand how the model makes decisions, we used LIME (Local Interpretable Model-agnostic Explanations). LIME shows which parts of an image are most important for the model's prediction by slightly changing the image and observing how the prediction changes.



Results

Stats after 12 epochs:

Train Loss: 0.0136 | Train Accuracy: 99.74%

Val Loss: 0.0730 | Val Accuracy: 98.28%

Test Loss: 0.0751 | Test Accuracy: 98.4%

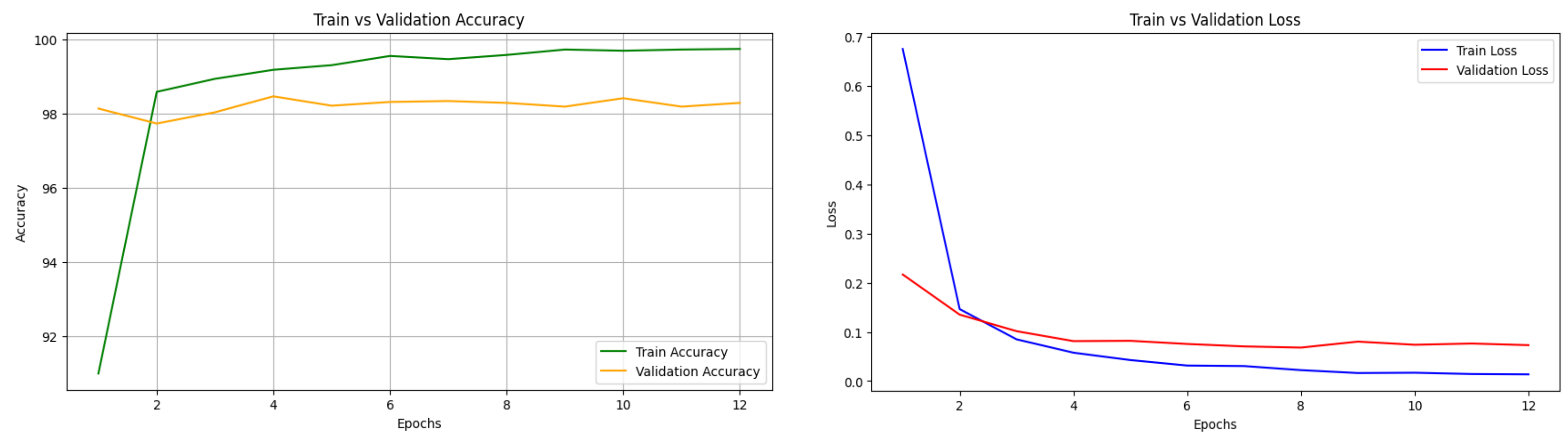
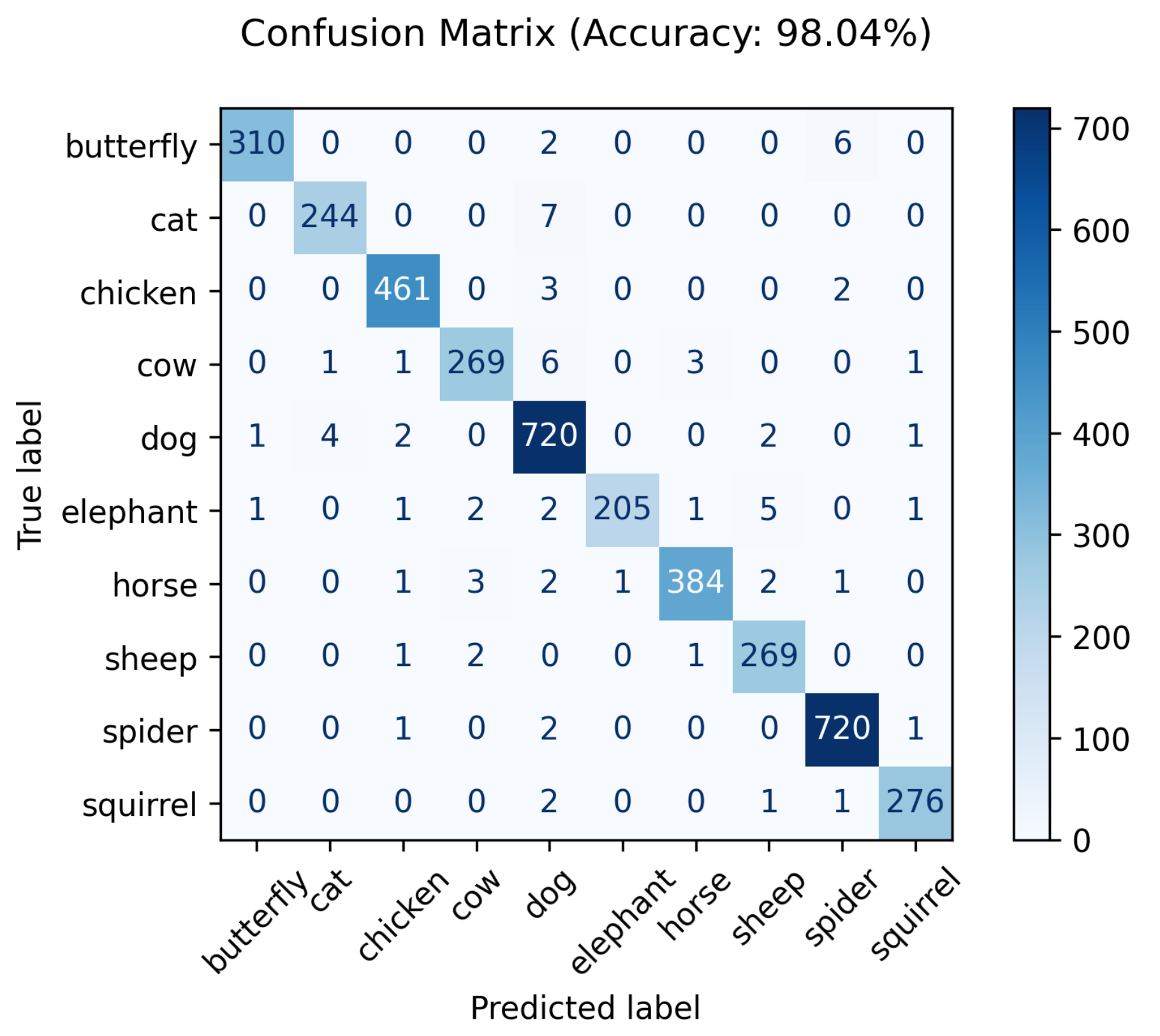


Figure: Training vs Validation Accuracy and Loss over Epochs



True label \ Predicted label	butterfly	cat	chicken	cow	dog	elephant	horse	sheep	spider	squirrel
butterfly	310	0	0	0	2	0	0	0	6	0
cat	0	244	0	0	7	0	0	0	0	0
chicken	0	0	461	0	3	0	0	0	2	0
cow	0	1	1	269	6	0	3	0	0	1
dog	1	4	2	0	720	0	0	2	0	1
elephant	1	0	1	2	2	205	1	5	0	1
horse	0	0	1	3	2	1	384	2	1	0
sheep	0	0	1	2	0	0	1	269	0	0
spider	0	0	1	0	2	0	0	0	720	1
squirrel	0	0	0	0	2	0	0	1	1	276

Figure: Confusion Matrix

Conclusion & Future Directions

Conclusion

The proposed Vision Transformer model achieved 98.41% accuracy in wildlife species classification, surpassing conventional CNN-based architectures such as ResNet-50 and EfficientNet-B4. The self-attention mechanism effectively extracted discriminative features without requiring manual feature engineering, as demonstrated by high precision (98.12%)and recall (98.35%) scores. Although hardware limitations restricted batch sizes and training speed, the results confirm ViT's strong potential for visual recognition tasks in ecological applications.

Future research directions should investigate:

Hybrid ViT-CNN architectures to combine strengths of both approaches

Advanced data augmentation techniques to handle com-plex backgrounds

Class-balancing strategies to improve performance on minority classes

These results suggest that transformer-based architectures can become the new standard for animal classification tasks, particularly when explainability and accuracy are crucial.

References

1. G. Horn, "Animals-10 dataset," Kaggle, 2021, https://www.kaggle.com/datasets/alessiocorrado99/animals10.
2. VIT journal: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,
3. M. J. Tabak, M. S. Norouzzadeh, and D. W. Wolfson, "Machine learning for camera trap images," Ecology and Evolution, vol. 9, pp. 2324–2336, 2019.