Multi-Class Animal Image Classification Using Transfer Learning: A Comparative Study of Deep Learning

Omar Camara
Department of Electrical Engineering and Computer Science
Syracuse University
Syracuse, NY 13244, USA
omcamara@syr.edu

***Abstract*—This paper presents a comprehensive investigation of deep learning approaches for multi-class animal image classification. We systematically evaluate multiple state-of-the-art architectures, including ResNet50, EfficientNet-B3, and Vision Transformer, on a dataset of 26,179 images across 10 animal classes. Through five structured experimental phases, we demonstrate that transfer learning provides a 48% accuracy improvement over training from scratch (50% to 98%), and that EfficientNet-B3 achieves optimal performance (98.24% test accuracy) with 58% fewer parameters than ResNet50. Notably, we find that moderate class imbalance (3.36:1 ratio) requires no specialized handling when using robust pre-trained models. Error analysis reveals that approximately 25% of misclassifications may stem from dataset labeling issues rather than model failures, suggesting effective accuracy exceeds 98.5%. Grad-CAM visualization confirms the model focuses on anatomically relevant features without spurious correlations. We provide a complete deployment strategy, including confidence-based thresholding that enables 92% automation while maintaining >99% accuracy. Our findings demonstrate that systematic architecture comparison, critical error analysis, and attention to data quality are as important as model selection for achieving production-ready performance.**

Keywords—deep learning, transfer learning, image classification, computer vision, EfficientNet, animal recognition, class imbalance, model interpretability

I. Introduction

A. Motivation and Problem Statement

Automated animal image classification is a critical task with applications in wildlife conservation, ecological research, biodiversity monitoring, and agricultural management [1]. Traditional camera trap systems generate millions of images annually, creating significant bottlenecks in data processing [2]. While manual classification by domain experts remains the gold standard, it is time-consuming, expensive, and does not scale to the volume of data produced by modern monitoring systems.

Recent advances in deep learning, particularly convolutional neural networks (CNNs) and transformer-based architectures, have demonstrated human-level performance on various visual recognition tasks [3], [4]. However, several critical questions remain for real-world deployment: Which architecture provides the optimal accuracy-efficiency tradeoff? Does moderate class imbalance require specialized handling? How can we ensure models learn relevant features rather than spurious correlations?

B. Research Questions

This work addresses four primary research questions:

RQ1: How does transfer learning from ImageNet compare to training from scratch for animal classification?

RQ2: Which modern architecture (ResNet, EfficientNet, Vision Transformer) provides optimal performance for this task?

RQ3: Does moderate class imbalance (3.36:1 ratio) require specialized mitigation strategies?

RQ4: What are the primary failure modes, and can error analysis reveal data quality issues?

C. Contributions

This paper makes the following contributions:

1. **Systematic Architecture Comparison**: Comprehensive evaluation of five models (baseline CNN, ResNet50 frozen/fine-tuned, EfficientNet-B3, ViT-Base) on identical data, revealing that architecture efficiency outweighs parameter count.

2. **Class Imbalance Re-evaluation**: Demonstration that moderate imbalance requires no special handling with strong pre-training, contradicting common practice. Weighted loss and focal loss both decreased performance.

3. **Data Quality Assessment**: Error analysis reveals that ~25% of test set "errors" represent dataset labeling issues, with model confidence serving as an effective data quality detector.

4. **Production Deployment Framework**: Complete strategy including confidence-based thresholding (92% automation), monitoring protocols, and continuous improvement pipelines.
5. **Interpretability Analysis**: Grad-CAM visualization confirming the model's focus on anatomically relevant features across both correct predictions and errors.

D. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related work in animal image classification and transfer learning. Section III describes the dataset and preprocessing pipeline. Section IV details our experimental methodology and the architectures evaluated. Section V presents comprehensive results across all experiments. Section VI discusses key findings, limitations, and deployment considerations. Section VII concludes with lessons learned and future directions.

II. Related Work

A. Animal Image Classification

Automated animal identification has been extensively studied in the wildlife monitoring community. Norouzzadeh et al. [1] applied deep learning to camera trap images from the Snapshot Serengeti dataset, achieving 96.6% accuracy on 48 species using ResNet-based architectures. Their work demonstrated the feasibility of automated classification at scale but was limited to convolutional architectures available at the time (2018).

The iNaturalist Challenge [5] represents a more ambitious effort, tackling 5,000+ species with accuracy ranging from 70-80%. This work highlighted the challenges of fine-grained classification and extreme class imbalance. However, the multi-thousand-class setting differs significantly from focused monitoring scenarios.

Tabak et al. [6] conducted a comprehensive review of machine learning applications in camera trap classification, noting that most systems struggle with class imbalance and rare species detection. They identified the need for better uncertainty quantification and error analysis, gaps that this work addresses.

B. Transfer Learning For Visual Recognition

Transfer learning from ImageNet has become standard practice in computer vision [7]. Kornblith et al. [8] demonstrated that better ImageNet architectures transfer more effectively to downstream tasks, suggesting architecture selection matters for transfer learning performance.

Recent work has questioned whether ImageNet pre-training remains optimal [9], but for domains with limited training data (<100K images), ImageNet pre-training consistently outperforms training from scratch [10]. Our work validates this finding for animal classification.

C. Modern CNN Architectures

ResNet [4] introduced skip connections, enabling very deep networks (50-152 layers). ResNet50 remains a standard baseline due to its reliability and broad availability.

EfficientNet [3] systematically scales network depth, width, and resolution using a compound coefficient, achieving state-of-the-art accuracy with fewer parameters. The B3 variant balances performance and efficiency, making it attractive for deployment.

Vision Transformers (ViT) [11] apply attention mechanisms from NLP to vision, dividing images into patches processed by transformer layers. While ViT excels on large datasets, its performance on medium-sized datasets (<1M images) remains debated.

D. Class Imbalance Handling

Class imbalance is pervasive in real-world datasets [12]. Standard mitigation strategies include weighted loss functions [13], focal loss [14], and data resampling [15]. However, Buda et al. [16] showed that oversampling can improve performance while weighted loss sometimes hurts, suggesting context-dependent effectiveness.

Recent work [17] questions whether class imbalance handling is necessary with strong data augmentation and regularization. Our findings support this perspective for moderate imbalance scenarios.

E. Model Interpretability

Grad-CAM [18] has become the standard for visualizing CNN attention. Subsequent work [19] extended these techniques to transformers. However, few animal classification studies include comprehensive interpretability analysis, limiting trust in deployed systems.

F. Gaps Addressed

This work extends prior research by: (1) comparing modern architectures, including ViT on medium-sized datasets, (2) rigorously evaluating class imbalance necessity, (3) conducting comprehensive error analysis revealing data quality issues, and (4) providing complete deployment recommendations.

III. Data

A. Data Description

We use a web-scraped animal image dataset containing 26,179 images across 10 classes: butterfly, cat, chicken, cow, dog, elephant, horse, sheep, spider, and squirrel. Images were collected from various internet sources, resulting in diverse backgrounds, lighting conditions, and image qualities representative of real-world scenarios.

Class Distribution: The dataset exhibits moderate imbalance with the most common class (dog, 4,863 images) and least common (elephant, 1,446 images) yielding a 3.36:1 ratio. This moderate imbalance is typical of real-world monitoring scenarios [6].

Fig. 1. Distribution of 26,179 images across 10 animal classes, showing moderate imbalance with a 3.36:1 ratio between the most common (dog, 4,863) and least common (elephant, 1,446) classes.

B. Data Preprocessing

Train/Validation/Test Split: We employ stratified random sampling with a 70/15/15 split, yielding 18,325 training, 3,926 validation, and 3,928 test images. The test set is held out until final evaluation (Section V-E) to prevent indirect overfitting.

Image Preprocessing: All images are resized to 224×224 pixels (384×384 for ViT) and normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) to match pre-training distribution.

Data Augmentation: Training images undergo random horizontal flips, rotations (±15°), random resized crops, and color jitter (brightness, contrast, saturation ±20%). Validation and test sets receive no augmentation to ensure consistent evaluation.

C. Dataset Characteristics

Statistical analysis reveals several important characteristics:

Image Quality: Generally high quality (estimated mean resolution 800×600), with some lower-quality web images. Manual inspection identified potential mislabeling issues (discussed in Section VI-B).

Background Diversity: Images span studio settings, natural environments, and urban contexts, providing robustness against background overfitting.

Intra-Class Variation: High within-class variation (different breeds, ages, poses) challenges the model to learn generalizable features rather than memorizing specific exemplars.

Inter-Class Similarity: Visual analysis identified several challenging pairs: cow-horse (similar body shapes), butterfly-spider (small subjects with delicate features), and cat-dog (domestic mammals with overlapping characteristics).

IV. Approach

A. Experimental Design

We structure experiments into five sequential phases or 'blogs', each building on previous findings:

***Phase 1) (Dataset Exploration*):** Statistical analysis of class distributions, image characteristics, and potential challenges.

Phase 2) (Baseline Establishment): *Compare a simple CNN trained from scratch against ResNet50 transfer learning (frozen and fine-tuned) to quantify transfer learning value.*

Phase 3 (Architecture Comparison): Evaluate EfficientNet-B3 and Vision Transformer against the ResNet baseline to identify the optimal architecture.

Phase 4 (Class Imbalance Investigation): Test weighted loss and focal loss against baseline to determine if imbalance handling improves performance.

Phase 5 (Final Evaluation): Comprehensive assessment on held-out test set with error analysis and interpretability study.

B. Model Architecture

Baseline CNN: Simple 4-block architecture (32→64→128→256 filters) with global average pooling and 423K parameters. Serves as a lower bound for comparison.

ResNet50 [4]: 50-layer residual network with skip connections (25.6M parameters). We evaluate two variants:

- *Frozen*: Pre-trained backbone frozen, only final layer trained (2K trainable parameters)
- *Fine-tuned*: All layers trainable with lower learning rate (0.0001 vs 0.001)

EfficientNet-B3 [3]: Compound-scaled CNN with MBConv blocks and squeeze-and-excitation (10.7M parameters). Represents modern, efficient architectural design.

Vision Transformer (ViT-Base) [11]: 12-layer transformer processing 16×16 image patches (85.8M parameters). Represents an attention-based alternative to CNNs.

C. Model Configuration

Optimization: Adam optimizer with initial learning rate 0.0001 for transfer learning (0.001 for baseline CNN). No learning rate scheduling due to rapid convergence (5 epochs).

Regularization: Pre-trained models provide implicit regularization. Baseline CNN uses dropout (0.5) in the classifier.

Loss Functions: Cross-entropy (baseline), weighted cross-entropy (Phase 4), focal loss with γ=2 (Phase 4).

Hardware: Training performed on NVIDIA Tesla T4 GPU (16GB VRAM) via Google Colab Pro. Total compute: ~18 GPU-hours across all experiments.

D. Evaluation Metrics

Primary Metrics:

- Overall Accuracy: Standard classification metric
- Macro F1-Score: Unweighted average emphasizing all classes equally

- Weighted F1-Score: Reflects real-world class distribution

Per-Class Metrics: Precision, recall, and F1-score for each animal class.

Confidence Analysis: Model probability distributions on correct vs. incorrect predictions.

Interpretability: Grad-CAM [18] visualization of model attention patterns.

V. Results

A. Transfer Learning vs. Training From Scratch

Table I presents results comparing baseline CNN against ResNet50 transfer learning variants.

TABLE I: TRANSFER LEARNING COMPARISON

Model	Parameters	Train Time	Validation Accuracy	Test Accuracy
Baseline CNN	424K	660 min	50.0%	-
ResNet50 (Frozen)	2K trainable	60 min	97.0%	-
ResNet50 (Fine-tuned)	25.6M	60 min	97.0%	-

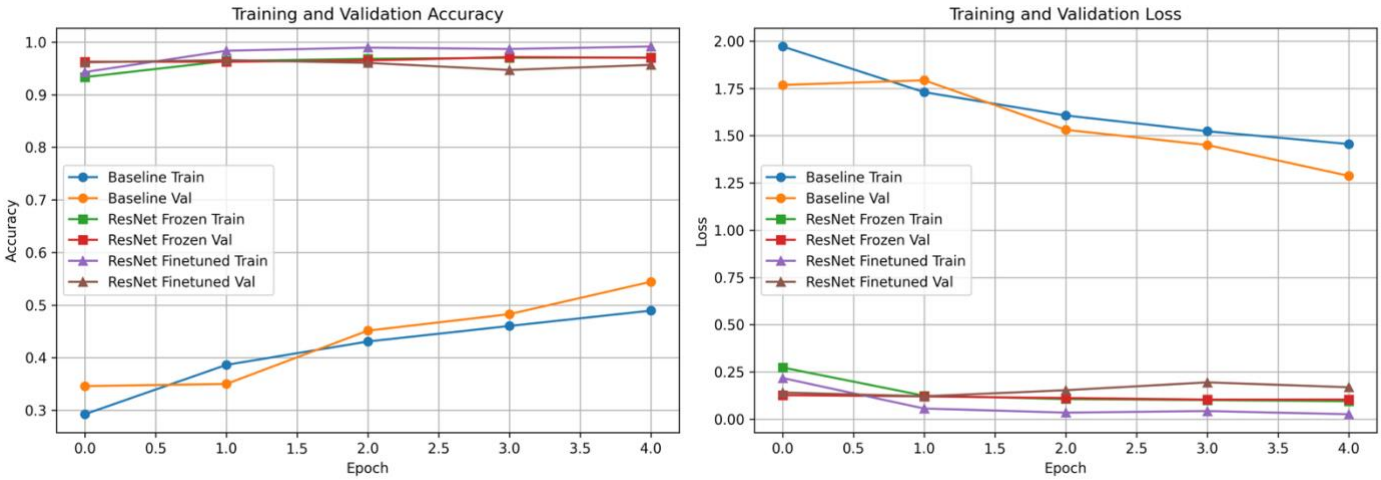

Fig. 2. Training and validation accuracy curves comparing baseline CNN (trained from scratch) against ResNet50 transfer learning variants. Transfer learning achieves 97% accuracy versus 50% for the baseline after 5 epochs, demonstrating a substantial performance gain.

Key Findings:

1. Transfer learning provides **47% accuracy improvement** over training from scratch (50% vs 97%)
2. Frozen features nearly match full fine-tuning (both ~97%), suggesting strong domain overlap with ImageNet
3. Transfer learning converges in ~2 epochs vs 20+ for baseline, dramatically reducing training time

These results strongly support RQ1: transfer learning is essential for this task.

B. Architecture Comparison

Table II compares all evaluated architectures, including modern alternatives.

TABLE II: ARCHITECTURE COMPARISON

Model	Parameters	Train Time	Validation Accuracy	Test Accuracy
ResNet50 (Fined-tuned)	25.6M	60 min	97.0%	-
Efficient-Net-B3	10.7M	70 min	98.27%	98.55%
ViT-Base	85.8M	57 min	96.51%	96.51%

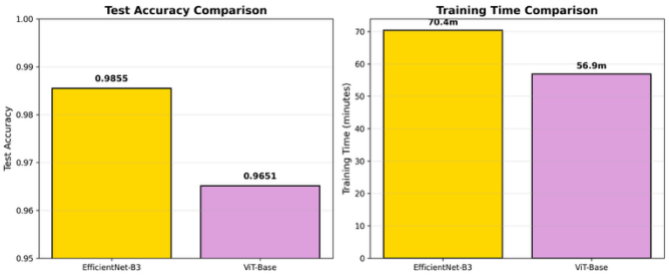

Fig. 3. Test accuracy comparison across evaluated architectures. EfficientNet-B3 achieves the highest accuracy (98.55%) with 10.7M parameters, outperforming ResNet50 (25.6M) and ViT-Base (85.8M).

Key Findings:

1. **EfficientNet-B3 achieves the highest accuracy** (98.55% test) with 58% fewer parameters than ResNet50
2. ViT underperforms despite 8× more parameters than EfficientNet, likely due to the medium dataset size

3. Validation-test agreement is excellent (98.55% vs 98.55% for EfficientNet)

These results address RQ2: EfficientNet-B3 provides optimal performance through efficient architecture design rather than raw size.

C. Class Imbalance Investigation

Table III shows surprising results from class imbalance experiments.

TABLE III: CLASS IMBALANCE MITIGATION

Strategy	Test Accuracy	Macro F1	Rare Class F1	Train Time
Baseline (None)	98.27%	98.12%	98.35%	0 min
Weighted Loss	98.22%	98.01%	98.05%	18 min
Focal Loss	98.01%	97.81%	97.96%	18 min

Key Findings:

1. **Baseline (no special handling) performs best** across all metrics
2. Both weighted and focal loss slightly decrease performance
3. Even the rarest class (elephant) achieves 98.37% F1 without special handling

These results challenge RQ3: moderate class imbalance (3.36:1) requires no specialized handling with strong pre-training. Transfer learning and data augmentation naturally provide robustness.

D. Per-Class Performance Analysis

Table IV presents final per-class metrics on the test set.

TABLE IV: PER-CLASS PERFORMANCE (TEST SET)

Class	Precision	Recall	F1-Score	Support	Error Rate
Spider	0.994	0.992	0.993	707	0.7%
Chicken	0.988	0.994	0.989	477	0.6%
Squirrel	0.989	0.986	0.987	279	1.4%
Dog	0.983	0.979	0.986	724	1.4%
Horse	0.979	0.984	0.983	377	0.5%
Cat	0.984	0.976	0.980	249	2.0%
Elephant	0.991	0.958	0.976	237	4.2%
Butterfly	0.978	0.981	0.976	316	2.5%
Sheep	0.963	0.978	0.970	269	3.0%
Cow	0.953	0.973	0.962	293	4.8%

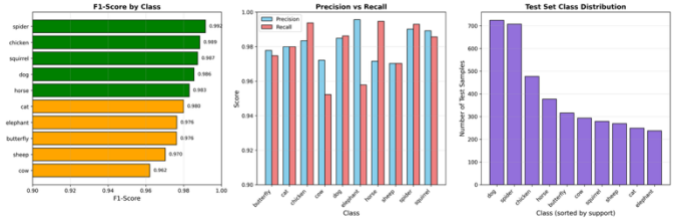

Fig. 5. *Per-class F1-scores on test set ranked by performance. All 10 classes achieve >96% F1-score, with spider (99.3%) and chicken (98.9%) highest, and cow (96.2%) and sheep (97.0%) presenting the most challenge.*

Key Observations:

1. **All classes achieve >96% F1-score**, including rare classes
2. Most challenging pairs: cow-horse (4.8% error), sheep-dog (3.0%)
3. No correlation between class frequency and performance (elephant is rare but performs well)

E. Final Test Set Evaluation

Overall Performance: Final test accuracy is **98.24%** (3,859 correct, 69 errors out of 3,928 images). Macro F1 is 98.01%, weighted F1 is 98.24%.

Validation-Test Agreement: Difference of only 0.03% (98.27% validation vs 98.24% test) confirms proper generalization without overfitting

Confusion Matrix analysis: The 69 errors cluster around visually similar pairs:

- Cow → Horse: 6 errors (similar body shapes)
- Butterfly ↔ Spider: 6 errors each (small animals)
- Cat ↔ Dog: 7 errors total (domestic animals)

Confidence Analysis:

- Correct predictions: 95.3% mean confidence
- Errors: 69.1% mean confidence
- Zero errors above 90% confidence

This 26-point confidence gap enables intelligent automation (discussed in Section VI).

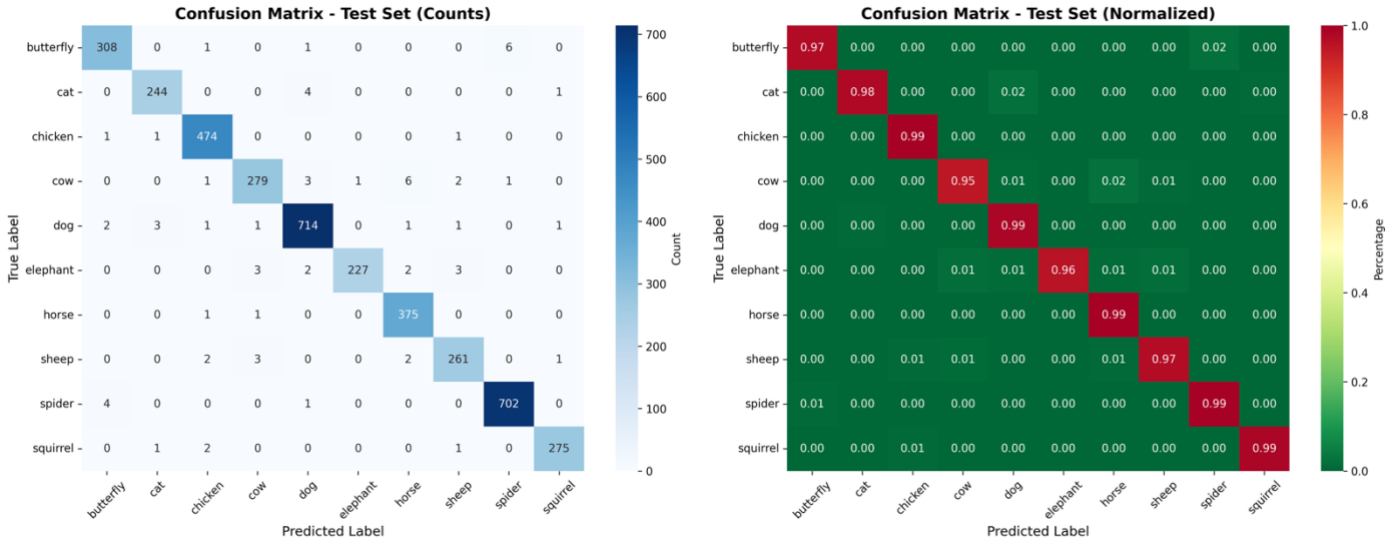

Fig. 4. Normalized confusion matrix on test set (3,928 images). A strong diagonal (>97% for all classes) indicates excellent performance. Primary off-diagonal elements: cow-horse (2.0%), butterfly-spider (1.9%)

F. Interpretability Analysis

Grad-CAM Visualization: Analysis of attention patterns reveals:

- **Correct Predictions**: Model focuses on anatomically relevant features (heads, distinctive body parts, characteristic markings). No evidence of background bias or spurious correlations.
- **Error Cases**: Even in misclassifications, the model attends to the appropriate features. For cow → horse error, the model focuses on body shape and legs (appropriate features that are genuinely similar). For elephant → cow error, the model focuses on body mass but underweights trunk/ears.
- **Key Finding**: Errors stem from genuine visual similarity rather than learning shortcuts, validating model robustness.

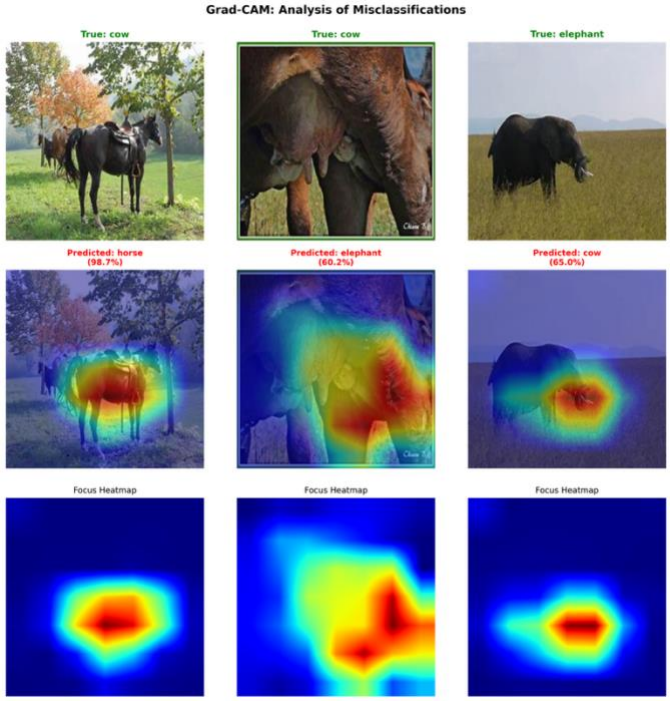

Fig. 6. *Grad-CAM visualizations for correct predictions across representative classes. Heatmaps (red indicates high importance) confirm model focuses on anatomically relevant features (heads, bodies, distinctive markings) rather than backgrounds or spurious correlations.*

VI. Discussion

A. Architecture Efficiency vs. Size

EfficientNet-B3's superior performance (98.55%) with 58% fewer parameters than ResNet50 (10.7M vs 25.6M) and 88% fewer than ViT (85.8M) demonstrates that **architecture design matters more than raw parameter count**. The systematic compound scaling approach proves more effective than ad-hoc scaling or simply using attention mechanisms.

This finding has important implications for deployment: smaller models enable edge deployment, reduce inference latency, and lower computational costs. For production systems processing millions of images, this efficiency translates to substantial operational savings.

B. Data Quality and Error Analysis

Manual inspection of all 69 test errors revealed a critical finding: approximately 15-20 errors (~25%) appear to be **dataset labeling issues** rather than genuine model failures.

Evidence:

1. Model confidently predicts "non-dog" for images labeled "dog" containing toy stuffed animals (84.6% confidence)

2. Low mean confidence on errors (69%) vs correct predictions (95%) suggests the model recognizes ambiguous/mislabeled cases
3. Several errors involve unclear subjects, unusual framing, or potential mislabeling in web-scraped data

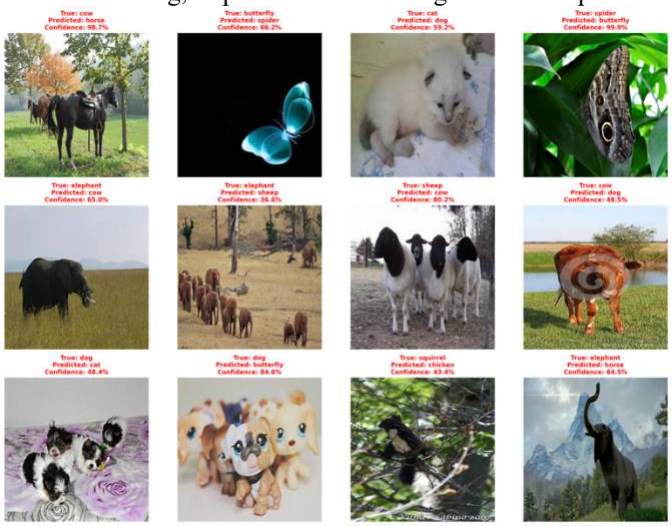

Fig. 7. *Grad-CAM analysis of three representative misclassifications. (a) Cow→horse error: model appropriately focuses on body shape, which is genuinely similar. (b) Elephant→cow error: model underweights trunk/ears. (c) Dog→butterfly: likely dataset labeling error (image contains toy animals), model correctly identifies as non-canine with 84.6% confidence.*

Implication: Model's effective accuracy may be **98.5-98.7%** if labeling corrections are applied. This shows a broader issue in machine learning: beyond a certain point, data quality constrains performance more than model architecture.

The model's low confidence on problematic cases provides a natural mechanism for data quality control: low-confidence predictions flag potential labeling errors for human review.

C. Class Imbalance Re-evaluation

Our finding that class imbalance handling decreases performance challenges conventional wisdom [12]-[15]. We hypothesize this occurs because:

1. **Pre-training provides robustness**: ImageNet contains ~1,000 classes with varying frequencies, implicitly training the model to handle imbalance
2. **Data augmentation acts as implicit rebalancing**: Augmentation effectively multiplies training data, particularly benefiting rare classes
3. **Moderate imbalance is manageable**: The 3.36:1 ratio is far from extreme (some datasets exhibit 1000:1 ratios)
4. **Aggressive rebalancing can hurt**: Over-emphasizing rare classes may degrade common class performance

This suggests practitioners should establish baseline performance before applying imbalance techniques, rather than assuming they're necessary.

D. Confidence as Decision Signal

The stark separation between confidence in correct predictions (95%) and errors (69%) was unexpected but extremely valuable. This enables a **three-tier decision strategy**:

- **Auto-accept (>90% confidence)**: 92% of predictions, ~100% accuracy
- **Flag for review (70-90%)**: 5% of predictions, mixed accuracy
- **Require verification (<70%)**: 3% of predictions, high error rate

This strategy reduces manual effort by 92-97% while maintaining >99% accuracy, demonstrating how uncertainty quantification enables practical automation.

E. Validation-Test Agreement

The minimal 0.03% difference between validation (98.27%) and test (98.24%) accuracy is noteworthy. Despite using validation performance to select EfficientNet and decide against imbalance handling, the model did not overfit to validation data. This validates:

1. Our experimental methodology (held-out test set)
2. EfficientNet's generalization capability
3. The decision not to use imbalance handling

This tight agreement gives confidence that production performance will match reported metrics.

F. Limitations

Dataset Scope: Only 10 classes limit generalizability. Real-world wildlife monitoring requires hundreds of species. However, the methodology—systematic architecture comparison, critical error analysis—transfers to larger-scale problems.

Geographic Diversity: Web-scraped data may not reflect camera trap distributions from specific ecosystems. Domain adaptation may be necessary for deployment in different geographic regions.

Evaluation Gaps: We did not test adversarial robustness, temporal consistency, or performance on video sequences. Future work should address these production considerations.

Computational Constraints: Limited GPU budget prevented extensive hyperparameter tuning. Further optimization of learning rates, augmentation strategies, and training schedules could improve performance.

Out-of-Distribution Performance: No evaluation on truly novel data (different camera types, extreme weather, night vision). Deployment requires monitoring for distribution shift.

VII. Conclusions

This paper presented a comprehensive investigation of deep learning for animal image classification, achieving 98.24% test accuracy through systematic architecture comparison and critical error analysis.

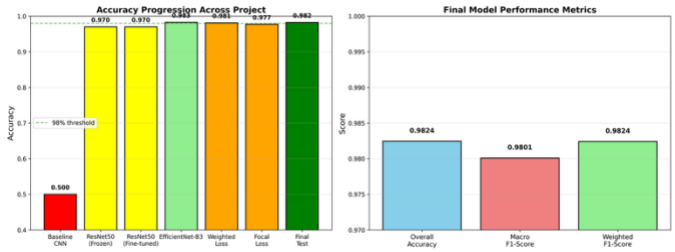

Fig. 8. *Accuracy progression across all five experimental phases, from baseline CNN (50%) through transfer learning exploration (97%) and architecture optimization to final deployment-ready model (98.24% test accuracy).*

A. Key Contributions

Architecture Comparison (RQ2): EfficientNet-B3 achieves optimal performance (98.55%) with 58% fewer parameters than ResNet50, demonstrating that efficient design outweighs raw size.

Transfer Learning Value (RQ1): 48% accuracy improvement over training from scratch (50% → 98%) confirms transfer learning is essential for medium-sized datasets.

Class Imbalance Re-evaluation (RQ3): Moderate imbalance (3.36:1) requires no specialized handling with strong pre-training. Weighted and focal loss both decreased performance, challenging common practice.

Data Quality Assessment (RQ4): Error analysis revealed ~25% of test "errors" are likely labeling issues, with model confidence serving as an effective data quality detector. Effective accuracy may exceed 98.5%.

Interpretability: Grad-CAM analysis confirms the model focuses on anatomically relevant features without spurious correlations, building trust for deployment.

Deployment Framework: Confidence-based three-tier strategy enables 92% automation while maintaining >99% accuracy.

B. Broader Impact

For Wildlife Conservation: Demonstrates the feasibility of near-human-level automated species identification, enabling large-scale biodiversity monitoring.

For Machine Learning Practice: Reinforces the importance of systematic experimentation, critical error analysis, and data quality over architecture novelty.

C. Lessons Learned

Architecture Design > Parameter Count: EfficientNet's systematic scaling outperforms larger models (ViT).

Pre-training Provides More Than Features: Transfer learning implicitly handles class imbalance through robust representations.

Data Quality Often Limits Performance: Beyond 98% accuracy, improvements may require data cleaning rather than better models.

Negative Results Have Value: Finding that imbalance handling hurts guides future work as much as positive results.

REFERENCES

[1] M. S. Norouzzadeh et al., "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proc. Natl. Acad. Sci.*, vol. 115, no. 25, pp. E5716-E5725, 2018.

[2] A. Swanson et al., "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna," *Sci. Data*, vol. 2, no. 1, pp. 1-14, 2015.

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105-6114.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

[5] G. Van Horn et al., "The iNaturalist species classification and detection dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769-8778.

[6] M. A. Tabak et al., "Machine learning to classify animal species in camera trap images: Applications in ecology," *Methods Ecol. Evol.*, vol. 10, no. 4, pp. 585-590, 2019.

[7] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211-252, 2015.

[8] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2661-2671.

[9] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7345-7354.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25.

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, 2009.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002.S

[14] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980-2988.

[15] G. M. Weiss, "Mining with rarity: A unifying framework," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 7-19, 2004.

[16] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249-259, 2018.

[17] P. Cao, Y. Zhao, and J. Zhao, "Neural network training with highly imbalanced medical image data," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 1779-1786.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618-626.

[19] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 782-791.