ISSN : 2455-9679
SJIF Impact Factor : 6.190



IJRTSM

INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"ARTIFICIAL INTELLIGENCE FOR CANINE SKIN-CONDITION IMAGE CLASSIFICATION: A CRITICAL REVIEW OF DEEP LEARNING, DATA QUALITY, EXPLAINABILITY, AND CLINICAL TRANSLATION"

Deepak Thakre [1], Dr. Dilip Solanki [2]

[1] P. G. Scholar, Institute of Advance Computing, SAGE University, Indore, Madhya Pradesh, India
[2] Professor, Institute of Advance Computing, SAGE University Indore, Madhya Pradesh, India

ABSTRACT

Canine dermatological disorders are challenging to classify from photographs because bacterial, fungal, and allergic conditions can produce overlapping visual signs, while coat colour, anatomical site, lighting, camera characteristics, and secondary infection alter lesion appearance. Artificial intelligence (AI), particularly convolutional neural networks (CNNs), offers a route to standardized image analysis, but reliable clinical translation requires more than high internal test accuracy. This narrative review examines the current evidence and methodological issues relevant to AI-based canine skin-condition classification, with emphasis on conventional image classification, transfer learning, multispectral imaging, feature fusion, dataset integrity, evaluation design, explainability, and external validation. Published work includes a multispectral-image study of three canine skin-disease categories that reported improved validation performance when predictions from conventional and multispectral images were combined, and a recent clinical-image study that trained EfficientNet models to identify visible lesion types such as erythema, lichenification, alopecia, and erosion/ulcer. These tasks are related but not equivalent: lesion recognition identifies visible morphology, whereas disease classification attempts to infer an underlying clinical category. The distinction affects label design, reference standards, and the claims that can be supported. Small datasets, augmentation leakage, uncertain labels, lack of dog-level identifiers, and limited external validation remain important risks. The review proposes a practical evaluation framework incorporating animal-level or source-group splitting, transparent baseline comparisons, class-wise metrics, calibration, error analysis, and clinician-led reference labelling. A hybrid EfficientNetB0–MobileNetV2 approach, as described in the associated project, is a plausible experimental hypothesis, but its superiority cannot be assumed without completed comparative testing. AI should presently be framed as a potential decision-support aid rather than a substitute for veterinary examination or diagnostic testing.

Key Words: *Canine dermatology; artificial intelligence; deep learning; convolutional neural networks; EfficientNet; MobileNetV2; transfer learning; explainable AI; veterinary imaging.*

I. INTRODUCTION

Skin disease is a common reason for veterinary assessment, yet visual presentation is rarely a one-to-one map from appearance to cause. Erythema, alopecia, scaling, crusting, papules, excoriation, and changes in skin texture may occur in more than one disorder. Bacterial and fungal disease can overlap visually, allergic disease may be complicated by secondary infection, and normal variation in coat, pigmentation, body site, and grooming can complicate interpretation.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Definitive diagnosis may require history, physical examination, cytology, culture, skin scraping, biopsy, or other targeted investigations. A photograph captures only part of this evidence.

Computer vision has become a useful research tool for extracting patterns from medical and veterinary images. CNNs learn hierarchical visual representations, and transfer learning allows a model pretrained on large image collections to be adapted to a smaller target dataset. In veterinary medicine, however, the evidence base is uneven across species and diagnostic modalities. A broad systematic review of veterinary deep learning identified applications across imaging and other clinical data, while also highlighting the importance of dataset diversity, interpretability, and expert involvement [1]. Canine dermatology is a particularly demanding setting because labels may describe either a visible lesion or a suspected underlying disease, and these are not interchangeable targets.

This review synthesizes selected peer-reviewed work and core methodological literature relevant to canine skin-image classification. It is a narrative critical review, not a PRISMA systematic review: the cited studies illustrate key directions and design challenges, but the manuscript does not claim exhaustive database coverage or a formal risk-of-bias assessment. The purpose is to identify what current evidence supports, what remains uncertain, and how future studies—including hybrid feature-fusion approaches—can be designed and reported more rigorously.

II. Review Scope and Approach

The review was informed by targeted searches for canine skin disease image classification, veterinary deep learning, multispectral imaging, lesion recognition, EfficientNet, MobileNetV2, and explainable AI. Priority was given to peer-reviewed veterinary studies, relevant review articles, and foundational architecture or explanation papers. The discussion distinguishes direct evidence from canine dermatology studies from methodological principles extrapolated from broader computer-vision research. Because no reproducible search log, duplicate screening, predefined eligibility criteria, or formal quality appraisal was completed, the paper should be submitted as a narrative review unless the authors conduct and document a systematic search.

III. Clinical and Data Challenges

3.1 Disease labels versus lesion labels

A central challenge is defining the prediction target. Disease-level labels such as bacterial dermatosis, fungal infection, or hypersensitivity/allergic dermatosis imply an underlying aetiology or clinical assessment. Lesion-level labels such as erythema, alopecia, lichenification, and erosion describe visible morphology. A model can potentially identify a visible lesion without determining its cause. Consequently, a study that performs well at lesion recognition should not automatically be interpreted as validating disease diagnosis. Label definitions should specify the clinical criteria, diagnostic tests used where appropriate, whether multiple labels are allowed, and how uncertain cases are handled.

3.2 Reference standards and annotation

Image labels are only as reliable as the process that generated them. Labels taken from web folders or dataset directories may not have the same evidential status as labels supported by a veterinary examination and relevant laboratory testing. Stronger datasets should document the reference standard for each category, the annotator's expertise, adjudication of disagreements, and inter-rater agreement. When a disease can only be distinguished using history or laboratory evidence, image-only classification should be presented as a prediction of the dataset label rather than a confirmed diagnosis.

3.3 Dataset size, imbalance, and acquisition variation

Small datasets can support pilot experiments but create high variance in model estimates and make overfitting more likely. Class imbalance can also conceal poor recognition of less frequent categories if evaluation relies on overall accuracy. Images collected from different cameras, clinics, coat colours, breeds, body sites, and lighting conditions may contain distribution shifts that are absent from a random split. Dataset documentation should therefore include counts by class, dog, source, clinic, device, and acquisition context where available. The absence of dog identifiers is especially limiting because multiple photographs of the same animal can make a nominally held-out test set less independent than it appears.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

IV. Deep-Learning Approaches for Canine Skin Images

4.1 CNNs and transfer learning

CNNs learn local visual features and progressively combine them into higher-level representations. Transfer learning reuses features learned from a large source dataset and adapts them to the target problem. This can be useful when veterinary image datasets are small, but the benefit depends on how closely the source features match the target domain and whether preprocessing is compatible with the pretrained model. The actual initialization, frozen layers, fine-tuning schedule, input size, and normalization must be reported. A fallback from pretrained weights to random initialization is a materially different experiment and should never be left undocumented.

4.2 EfficientNet and MobileNet families

EfficientNet introduced a compound scaling strategy that balances network depth, width, and input resolution [2]. MobileNetV2 uses inverted residual blocks and linear bottlenecks to reduce computational demand [3]. These architectures offer complementary engineering trade-offs: EfficientNet variants are often selected for a balance of representational capacity and cost, whereas MobileNetV2 is designed with resource-constrained inference in mind. Neither architecture is inherently optimal for canine dermatology. Their value must be measured on the target dataset using consistent splits, preprocessing, optimization, and checkpoint selection.

4.3 Feature fusion and ensemble strategies

Fusion combines information from multiple representations or models. In feature-level fusion, feature vectors from separate backbones are concatenated before classification. In decision-level fusion, probability outputs from separately trained models are combined. Multimodal fusion can also combine conventional photographs with other image modalities or clinical metadata. These strategies may improve robustness if the sources contain complementary information, but they also increase complexity and can overfit when the number of independent animals is small. Fusion should therefore be compared against single-model baselines, and any improvement should be evaluated alongside latency, memory use, calibration, and class-wise error patterns.

4.4 Multispectral imaging

Hwang and colleagues investigated canine skin-disease classification using conventional and multispectral images. Their study considered bacterial dermatosis, fungal infection, and hypersensitivity allergic dermatosis, with normal skin as the comparison category. The reported validation accuracy of the best single-modality models was 0.80 for bacterial dermatosis, 0.70 for fungal infection, and 0.82 for hypersensitivity allergic dermatosis; the best consensus models combining conventional and multispectral information reached 0.89, 0.87, and 0.87, respectively [4]. These results suggest that complementary image information can be useful, but they should be interpreted in light of the study's limited original image collection and augmentation strategy. Large numbers of augmented images do not create the same independent information as a large number of separately sampled dogs.

4.5 Recent lesion-level classification

Kang and colleagues reported CNN-based identification of four common canine skin lesion types—erythema, lichenification, alopecia, and erosion/ulcer—from clinical images. Four EfficientNet models were trained independently for the lesion categories, and all reported accuracies exceeded 90%; the alopecia model achieved 98.12% accuracy and an F1 score of 98.18%, while lichenification had the lowest reported sensitivity and F1 score among the four tasks [5]. The study is relevant evidence that common lesion morphology can be learned from clinical photographs. Its task is nevertheless lesion identification rather than direct classification of bacterial, fungal, and allergic aetiologies. It should not be treated as a like-for-like benchmark for disease-category classifiers.

V. Critical Comparison of Representative Approaches

Table 1. Selected approaches relevant to canine skin-image analysis. The studies address different prediction targets and evaluation settings; their reported scores should not be ranked as if they were measured on a common test set.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Approach / study	Prediction target	Strength	Key caveat
CNN transfer learning	Image or disease category	Reuses pretrained visual features; practical for limited data	Sensitive to domain shift and label quality
EfficientNet family	Single-image classification	Scalable accuracy–compute trade-off [2]	Must be compared on identical splits; no architecture guarantees clinical validity
MobileNetV2	Single-image classification	Designed for efficient inference [3]	Lower compute cost does not resolve data bias or ambiguous labels
Conventional + multispectral consensus [4]	Three disease categories vs normal	Reported higher validation performance than single-modality models	Small original image collection; augmentation and split design affect interpretation
EfficientNet lesion models [5]	Erythema, lichenification, alopecia, erosion/ulcer	Clinical image study with lesion-specific evaluation	Morphological recognition is not equivalent to aetiological diagnosis
Hybrid feature concatenation	Four project-defined categories	Can test whether backbones encode complementary information	Added complexity; improvement remains a hypothesis until directly measured

VI. Data Leakage and Evaluation Design

6.1 Duplicates and augmentation lineage

Exact-file hashes detect byte-identical files but miss resized, cropped, recompressed, colour-adjusted, or otherwise transformed versions of the same photograph. Perceptual hashes, image-embedding similarity, and source-image metadata can help identify related examples. Crucially, all augmented versions of one source image must remain within a single partition. If near-duplicates cross training and test sets, the model may be rewarded for recognizing image-specific details rather than learning generalizable disease patterns.

6.2 Splitting at the correct unit

The most defensible split unit is generally the animal, not the individual photograph, when multiple images per dog exist. Where dog identifiers are unavailable, a source-group split is a useful partial safeguard but cannot establish animal-level independence. If data come from multiple clinics, a clinic-held-out evaluation can provide additional evidence of transportability. The test set should be locked before model selection and used only for final evaluation. Validation data—not test data—should guide hyperparameter choices, threshold selection, early stopping, and checkpoint selection.

6.3 Metrics and uncertainty

Accuracy is easy to communicate but can hide class-specific failures. At minimum, studies should report per-class precision, recall/sensitivity, specificity where meaningful, F1 score, macro-F1, weighted-F1, and a confusion matrix. For probabilistic models, calibration measures and reliability plots can help assess whether predicted probabilities correspond to observed frequencies. Confidence intervals are important because small test sets produce uncertain estimates. ROC and precision–recall analyses should specify one-vs-rest or macro/micro aggregation and should not replace clinically interpretable error analysis.

6.4 Baselines and fair comparison

A proposed hybrid EfficientNetB0–MobileNetV2 system should be compared with EfficientNetB0 alone and MobileNetV2 alone under identical data partitions, augmentation, preprocessing, optimizer settings where feasible, and

ISSN : 2455-9679
SJIF Impact Factor : 6.190

checkpoint policy. Results should include performance differences and uncertainty, not just the highest point estimate. Model size, inference latency, memory requirements, and reproducibility also matter. If the hybrid model does not provide a reliable benefit, a smaller single backbone may be the more appropriate engineering choice.

VII. Explainability and Error Analysis

Grad-CAM uses gradients to produce a coarse map of image regions associated with a class score [6]. Such maps can support qualitative inspection: for example, whether predictions appear to rely on a lesion rather than a watermark, hand, collar, background, or image border. They do not prove that a model has learned a causal or clinically valid mechanism. Explanation maps should be examined across all classes and should include both correct and incorrect predictions. Error review should consider lesion size, image focus, lighting, body site, fur coverage, label uncertainty, and the clinical consequences of false negatives and false positives.

VIII. Clinical Translation, Ethics, and Deployment

An image classifier may eventually support preliminary triage, documentation, or consistency in lesion description. Before clinical use, however, developers must specify the intended user and use case. A research tool used by a veterinarian differs from an application that encourages pet owners to self-diagnose or treat an animal. Prospective evaluation should measure performance on independently collected cases, subgroup differences, calibration, usability, and whether access to predictions changes clinical decisions. Clear escalation advice and explicit statements of uncertainty are essential. The system should not recommend treatment solely from a photograph or imply that it replaces examination and diagnostic tests.

Data governance also matters. Image collection should follow applicable animal ethics and owner-consent requirements, and identifying metadata should be removed where possible. Data availability statements should accurately describe access restrictions. Code, preprocessing specifications, model weights where shareable, random seeds, software versions, and evaluation scripts improve reproducibility. Funding, conflicts of interest, author contributions, and ethical approvals should be disclosed according to the target journal's policies.

IX. Research Gaps and Recommended Agenda

- Build multi-centre datasets with animal-level identifiers, diverse breeds, coat colours, skin pigmentation, anatomical sites, devices, and disease severity.
- Use explicit diagnostic reference standards and record uncertainty, multiple concurrent conditions, and annotation disagreements.
- Separate lesion morphology recognition from disease-aetiology classification, and use multi-label outputs when clinically justified.
- Use animal-level or source-group splits and reserve an independent external test cohort from different clinics or acquisition settings.
- Compare simple baselines, single backbones, and fusion models under a common protocol; report uncertainty and computational cost.
- Evaluate calibration, subgroup performance, and clinically weighted errors in addition to accuracy and F1.
- Use Grad-CAM as one qualitative tool alongside blinded error review, not as proof of model reasoning.
- Conduct prospective workflow studies before making claims about improved clinical outcomes or decision support.

X. Relevance to a Hybrid EfficientNetB0–MobileNetV2 Framework

The associated project proposes to concatenate pooled features from EfficientNetB0 and MobileNetV2 and classify four dataset-defined categories: bacterial dermatosis, fungal infections, healthy skin, and hypersensitivity/allergic dermatosis. The supplied project description reports 442 readable images, near-duplicate pairs and label-review flags, and a planned group-aware split. These details identify a meaningful methodological case study, but they do not

ISSN : 2455-9679
SJIF Impact Factor : 6.190

establish model effectiveness. Because training and final evaluation have not been completed, the framework should be described as proposed rather than validated.

Before results are presented, the flagged images and cross-class near-duplicate pair should be reviewed; the class mapping and preprocessing should be verified; pretrained-weight initialization should be recorded; and checkpoint selection should use a consistent validation criterion. The hybrid system should then be compared fairly with both single-backbone baselines. Only after the pipeline is run should the authors report test metrics, confusion matrices, training curves, Grad-CAM examples, model size, and inference latency. The four class names should remain explicitly identified as project-defined labels unless their diagnostic validity is independently established.

XI. Conclusion

AI-based canine skin-image analysis is a promising but still methodologically demanding area. Published studies show that CNNs can recognize visible canine skin lesions and that combining complementary image information may improve selected validation outcomes. Yet disease classification from photographs remains constrained by overlapping signs, uncertain labels, small or non-independent datasets, and limited external validation. Reliable progress depends on clinically grounded annotation, leakage-resistant evaluation, fair baselines, transparent reporting, calibrated predictions, and prospective testing. Hybrid feature fusion is a testable design choice—not a guarantee of better performance. Until independent clinical validation supports a defined use case, such systems should be treated as research or decision-support tools and not substitutes for veterinary diagnosis.

REFERENCES

[1] Xiao S, Dhand NK, Wang Z, Hu K, Thomson PC, House JK, Khatkar MS. Review of applications of deep learning in veterinary diagnostics and animal health. Frontiers in Veterinary Science. 2025;12:1511522. https://doi.org/10.3389/fvets.2025.1511522

[2] Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research. 2019;97:6105–6114.

[3] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:4510–4520.

[4] Hwang S, Shin HK, Park JM, Kwon B, et al. Classification of dog skin diseases using deep learning with images captured from multispectral imaging device. Molecular & Cellular Toxicology. 2022;18:299–309. https://doi.org/10.1007/s13273-022-00249-7

[5] Kang SY, Kang YH, Kim HJ, Huh EA, Jeon JS, Kim HS, Kim MS, Tiwari A, Hwang CY. Artificial intelligence-based identification of common canine skin lesions from clinical images. Veterinary Dermatology. 2026;37(4):471–477. https://doi.org/10.1111/vde.70083

[6] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:618–626.

[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770–778.

[8] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009:248–255.