Classification of Pet Animals Skin Disease Using Vision Transformers

P. Gowsikraja, S. S. Samwilson, S. Joedinesh and K. Saran
Department of Computer Science and Design, Kongu Engineering College (Autonomous) Perundurai,
Erode - 638 060, Tamil Nadu, India

Keywords: Vision Transformers, Pet Skin Diseases, Deep Learning, Image Classification, Veterinary Diagnostics.

Abstract: The classification of pet skin diseases is becoming more and more necessary due to the increasing demand for early and accurate diagnostics in veterinary healthcare. Complex and diverse skin disease patterns of pets urge advanced models for reliable recognition and diagnosis. In this paper, we propose an approach based on Vision Transformers (ViTs) for the classification of skin diseases in pets. The proposed model would be capable of detecting and categorizing a number of skin diseases efficiently after processing images in such a way that detailed features crucial for accurate classification are captured. The model has been trained on a dataset of annotated images to better generalize and perform. These results will be of very important significance in the diagnosis of different skin diseases found in pets and will contribute a lot to veterinary dermatology while propagating modern deep learning techniques in improving diagnostic accuracy.

1 INTRODUCTION

It reflects a growing demand for veterinary care, especially dermatology, due to the skin diseases in pets that are very common. Therefore, these conditions should be diagnosed early and correctly classified so that proper treatment can be initiated, hence overall health of the animals. However, traditional methods used for diagnostics present an unsuitable case as disease manifestation variability is evident, even in front of seasoned veterinarians. Techniques based on deep learning are indeed very promising as they may well automate such diagnoses and bring improved accuracy into it. A new architecture in the computer vision domain, known as Vision Transformers or ViTs, was surprisingly effective for most classification tasks. This is due to its ability to model long-range dependencies in images. The paper is based on the application of ViTs for the classification of pet skin disease. It elaborates on how it applies, performs, and compares with other traditional models, such as CNNs. In this study, the aim is to design a robust model for distinguishing various skin diseases in pets and to improve diagnostic efficiency. Traditional techniques of diagnosis are mostly based on observation, and hence, vary because of the asymmetrical presentation of symptoms. This brings out the emerging importance of AI-based techniques to aid in support for diagnostic activities, especially for image classification. Deep learning architectures, especially of advanced form known as Vision Transformers, have shown a lot of promise in image analysis and classification that even includes medical images. We will apply the application of ViTs on a pet skin disease classification task and extend veterinary healthcare diagnostics from a set of images specially labeled for different kinds of skin diseases.

2 RELATED WORKS

Artificial intelligence, particularly deep learning, has made significant strides in medical image classification. Convolutional Neural Networks (CNNs) have long been the dominant architecture in image classification tasks, with numerous studies demonstrating their effectiveness in identifying human skin diseases such as melanoma and psoriasis. These models excel at extracting hierarchical features but can struggle when handling the complex patterns often present in pet skin diseases.

The Vision Transformer architecture offers an alternative to CNNs by treating images as a sequence of patches, which allows the model to retain information from across the entire image. Unlike CNNs, which focus on local feature extraction through convolutional filters, ViTs can process global image information more effectively. This characteristic

Gowsikraja, P., Samwilson, S. S., Joedinesh, S. and Saran, K.
Classification of Pet Animals Skin Disease Using Vision Transformers.
DOI: 10.5220/0013875300004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 1*, pages 888-893
ISBN: 978-989-758-777-1

makes them well-suited for tasks where the relationships between distant regions of an image are important, such as in the classification of skin diseases with varied presentations. While ViTs have shown promise in fields like object detection and medical imaging, their application in veterinary diagnostics remains relatively unexplored.

3 METHODOLOGY

3.1 Dataset Preparation

Table 1: Dataset Distribution by Disease Type.

Disease Type	Number of Images	Percentage (%)
Healthy Skin	200	12%
Fungal Infections	350	21%
Bacterial Dermatosis	300	18%
Hypersensitivity Allergic Dermatosis	350	21%
Other Conditions	100	6%
Total	1300	100

For this table 1, the dataset consists of images of three primary skin disease categories in dogs namely Fungal Infections, Hypersensitivity Allergic Dermatosis, and Bacterial Dermatosis and includes data of healthy skin to differentiate affected and healthier skins. These images were collected from veterinary clinics and medical repositories, ensuring that the dataset covers a variety of symptoms and conditions for each disease type.

- Fungal Infections: These include conditions caused by fungi, such as ringworm, characterized by scaly patches and hair loss.
- Hypersensitivity Allergic Dermatosis: This refers to allergic reactions leading to skin inflammation, itching, and swelling.
- Bacterial Dermatosis: These infections cause redness, lesions, and, in some cases, pus formation on the skin.

Each image in the dataset was manually labelled by veterinary professionals to ensure accurate and high-quality annotations. To prevent overfitting and improve model generalization, data augmentation techniques such as random rotations, scaling, and flips were applied. This ensured that the model would perform well across diverse scenarios.

The dataset was divided into three subsets: training (70%), validation (15%), and testing (15%), with the image resolution standardized to 224x224 pixels to fit the input size requirements of the Vision Transformer model.

3.2 Vision Transformer Model

This architecture based on ViT changes the processing of an image in comparison with the standard CNNs. The model does not rely on the usual convolutional layers of the CNN to extract features but instead functions by splitting the input image into non-overlapping small patches. These are usually 16 × 16 pixels in size; then flattened to a one-dimensional vector. It enables the model to take images as a sequence, just like the way NLP tasks are addressed in processing text. This embedding space allows these vectors to be projected once the image is divided into patches. In this context, the embedding becomes crucial for allowing the model to understand relationships among different patches. The transformer layers are the backbone of the Vision Transformer. These employ self-attention mechanisms, which gives more relevance to some patches with respect to the other patches, allowing the model to capture more patterns and the correlations between the local and the global features that exist in the images.

The multi-head attention mechanism in the transformer layers gives a deep, nuanced understanding of content through focusing differently at various places of the images at once. This is especially the case for disease classification problems, where sometimes slight differences in texture and structure can signify different skin diseases. The output of the transformer layers proceeds to a classification head, often a Multi-Layer Perceptron (MLP), to produce the final predictions for the categories for the skin disease. See the architecture of the vision transformer in Figure 1.

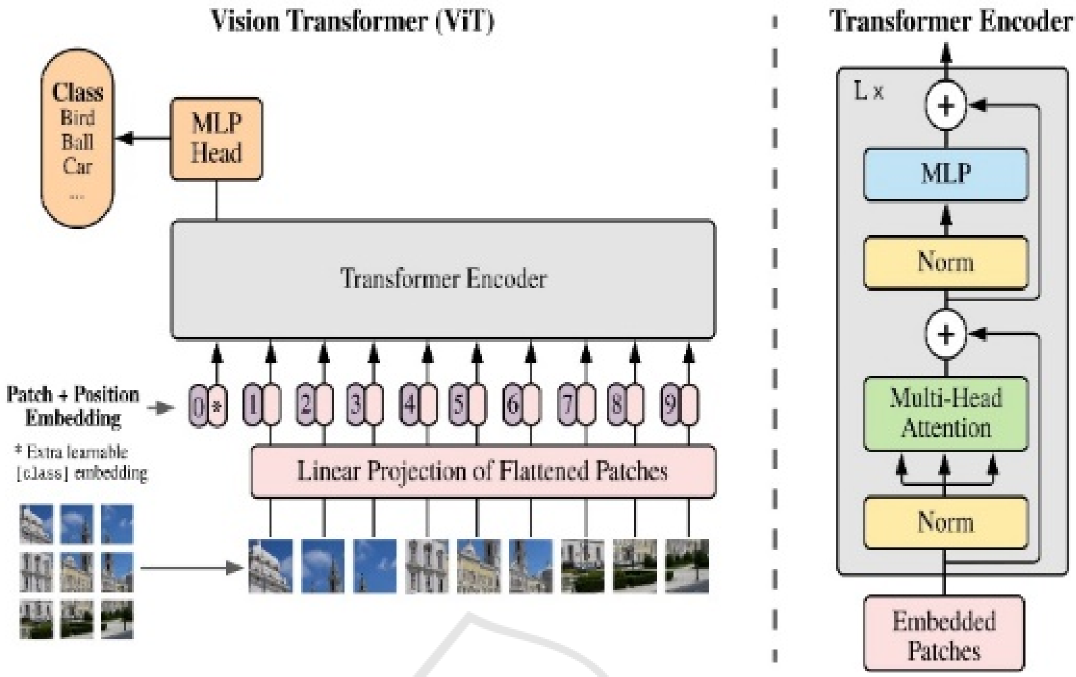

Figure 1: Vision Transformer Architecture.

3.3 Model Specifications

The Vision Transformer model for this study was configured with the following specifications:

- **Input Image Size**: 224×224 pixels. This size is commonly used to maintain a balance between computational efficiency and sufficient detail for accurate classification.
- **Patch Size**: 16×16 pixels. This patch size allows for capturing localized features while maintaining a manageable number of patches for processing.
- **Number of Layers**: 12 transformer layers. The depth of the model helps it learn complex representations.
- **Embedding Dimension**: 768. This high-dimensional space allows for a more expressive representation of the input data.
- **Attention Heads**: 12. Multiple attention heads enable the model to attend to different parts of the image, capturing various features effectively.
- **Classification Head**: Multi-Layer Perceptron (MLP), which processes the aggregated information from the transformer layers and outputs probabilities for each class of skin disease.

3.4 Flow Chart

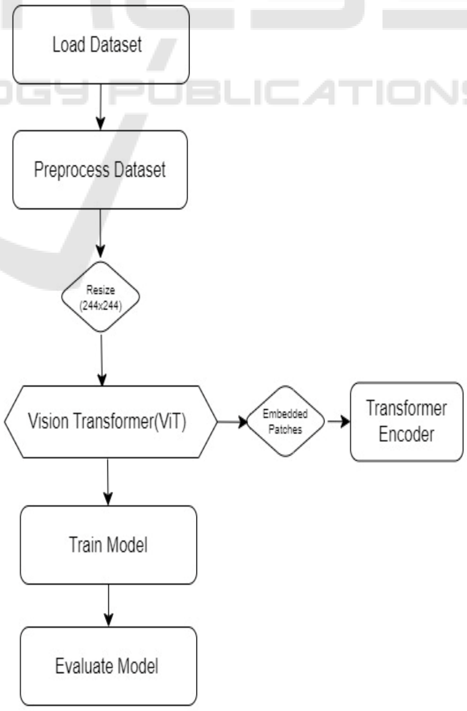

Figure 2: Flow Chart.

3.5 Training Process

The training process followed a stepwise procedure to the Vision Transformer, thus ensuring correct learning and proper generalization for unseen data. Using the Adam optimizer was pertinent, considering that it efficiently takes care of sparse gradients and adjusts learning rates in real time. The chosen learning rate was 0.0001, balanced to provide maximum convergence speed within stable training processes. Training has been performed in batches of 32 images; this is ensured to provide sufficient computation with regard to memory use. Batch size influences the generalization capability of the model; too large a batch size tends to generalize poorly while the opposite may lead to noisy updates. The loss function is cross-entropy that is well suited for multi-class classification problems. This function calculates the difference of class probabilities, which the model has predicted with the actual class labels, and these differences guide the model to reduce its prediction errors. It is trained for 50 epochs with the early stopping mechanisms, which mean that the training is stopped if the model fails to improve for a specified number of epochs on the validation set. This can prevent overfitting in the models, a typical problem that often arises in deep learning models when the dataset is relatively small.

The above random rotations, scaling, flipping, and colour augmentation are then used to build the model as robust. The model learns it to be exposed to a lot more scenarios than it would otherwise have been under if all of these were to have been learned on live data. Finally, there was transfer learning. Here the initialization is with the weights attained by pretraining on the ImageNet dataset. This approach now puts the knowledge learned from a large-scale dataset to good use: it now constitutes an extremely strong initialization for the model. It would allow the ViT model to adapt even better to the particular task of classification of skin disease, even by being compatible with a relatively small target dataset.

4 RESULT AND DISCUSSIONS

4.1 Performance Evaluation

Test set performance was conducted on the ViT model by focusing considerably on major three key performance metrics of evaluation-precision, recall, F1-score, and accuracy. The overall accuracy rate returned was 90.5%, showing how the model can classify images very well according to different categories of skin diseases. Besides this, all precision, recall, and F1-score of the model were more than 90% with respect to diagnostic accuracy, as well as consistency in finding accuracy with the right disease. The precision value refers to the accuracy with which the model will classify its positive instances without producing excessive false positives. It is very reliable for diagnostics. The recall score shows how good the model is in classifying true positives and, thus, avoiding a possible misclassification. The high F1-score combines precision and recall, indicating that the model has the strength to handle the complexity associated with variable presentations of the disease.

These results confirm reliability for models in real-veterinary and actual medical application fields where diagnoses are important as they have to be both precise yet speedy. The ViT Vision transformer can view full images and identify between two very similar conditions so that such specifics appear well distinct. In this success, therefore, deep learning models like ViT are most likely to change practice in the sense that they are going to make such a diagnosis. This is concerning areas such as dermatology. Table 2 represents the model performance metrics and figure 3 and 4 shows the model evaluation metrices and confusion matrix.

Table 2: Model Performance Metrics.

Metric	Value
Accuracy	90.5%
Precision	91.2%
Recall	90.8%
F1- Score	91.0%

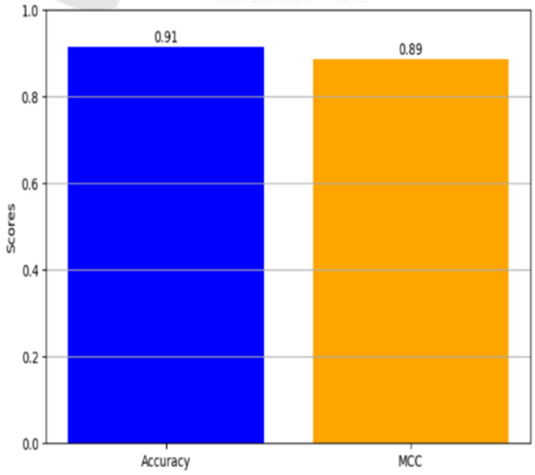

Figure 3: Model Evaluation Metrices.

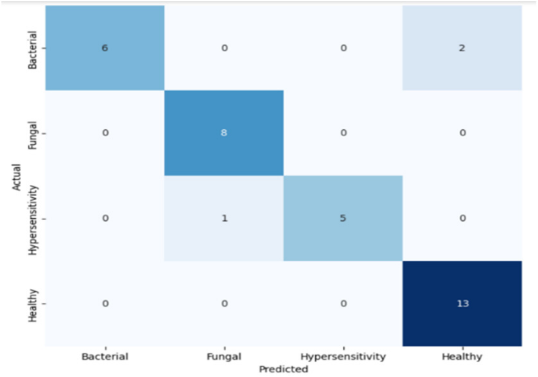

Figure 4: Confusion Matrix.

4.2 Discussion

The model proved to outperform on the different classes of disease. It can extract a very wide-ranging pattern in the entire image and proved to be way better than approaches that highly depend on these local features. Actually, it showed a really impressive performance between different skin conditions which look pretty similar, such as fungal infections versus bacterial infections. Still, the scope to enhance the performance further is still in place. Generalizing capabilities of this model will also be remarkably high if diversity, such as greater instances of uncommon skin diseases or even broader species than cats and dogs, of the dataset will be enhanced. This would ensure that more precise diagnoses occur in real veterinary practice as its capability to robustly deal with different scenarios it might face would be increased.

5 CONCLUSIONS

This research demonstrates the promising capabilities of the ViTs in accomplishing the task of skin disease classification in pets. The results show that it was found to be accurate under different conditions, hence really applicable to help veterinary diagnostics get better. Application of ViTs would therefore greatly enhance the accuracy as well as the speed of diagnosis, which, in a way, assists in better planning of treatment for pets, indirectly enhancing their quality of life. In particular, the ability to pull complex patterns and global features from images will be helpful in differentiating presentations of skin disease that may look very much alike. This feature of the ViTs will prove to be of utmost value in veterinary practice, thanks to the ability to make a diagnosis rapidly and precisely, therefore arresting the course of disease and allowing intervention sooner rather than later.

Add new data modalities, for example, clinical notes or owner-reported symptoms and treatment history, in order to have a holistic diagnostics approach that is going to stretch the model very far. A multimodal approach like this one would add visual context, beside the mere analysis of vision; it increases the accuracy of the classification more than just purely visual analysis of the model above. It will also require a dataset large and varied enough to cover different skin conditions and species of pets so that the model would be robust and generalizable in real-world veterinary settings.

In summary, the successful application of Vision Transformers in the present study suggests that these have the potential to revolutionize the face of veterinary diagnostics. With development in the future, technologies of AI are bound to change how veterinarians will handle diagnostics and treatment on balance, better outcomes for the pets and less stress among the owners.

REFERENCES

Bhavsar, S., & Mehendale, N. (2022). Deep Learning-Based Automatic System for Diagnosis and Classification of Skin Dermatoses. https://doi.org/10.21203/rs.3.rs- 2360579/v1

Gupta, P., & Gupta, S. (2022). Deep learning in medical image classification and object detection: A survey. International Journal of Image Processing and Pattern Recognition. https://doi.org/10.37628/ijippr.v8i2.846

Himel, G. M., Islam, Md. M., Al-Aff, Kh. A., Karim, S. I., & Sikder, Md. K. (2024). Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system. International Journal of Biomedical Imaging, 2024, 1–18. https://doi.org/10.1155/2024/3022192

Hwang, S., Shin, H. K., Park, J. M., Kwon, B., & Kang, M.-G. (2022). Classification of dog skin diseases using deep learning with images captured from Multispectral Imaging device. Molecular &amp; Cellular Toxicology, 18(3), 299–309. https://doi.org/10.1007/s13273-022-00249-7

Hyeon Ki Jeong 1 2, 1, 2, & Artificial intelligence (AI) has recently made great advances in image classification and malignancy prediction in the field of dermatology. However. (2022, August 23). Deep learning in dermatology: A systematic review of current approaches, outcomes, and limitations. JID Innovations. https://www.sciencedirect.com/science/article/pii/S2667026722000583

Jiang, Z., Gu, X., Chen, D., Zhang, M., & Xu, C. (2024). Deep learning-assisted multispectral imaging for early screening of skin diseases. Photodiagnosis and Photodynamic Therapy, 48, 104292. https://doi.org/10.1016/j.pdpdt.2024.104292

Mohsin Ali, M., Chandra Joshi, R., & Kishore Dutta, M. (2022). An automated and efficient deep learning-based classification of multiple skin disorders from skin lesion images. 2022 International Conference on Edge Computing and Applications (ICECAA), 1156–1161.https://doi.org/10.1109/icecaa55415.2022.9936097

Rathnayaka, R. M. N. A., Anuththara, K. G. S. N., Wickramasinghe, R. J. P., Gimhana, P. S., Weerasinghe, L., & Wimalaratne, G. (2022). Intelligent system for skin disease detection of dogs with ontology based clinical information extraction. 2022 IEEE 13th Annual Ubiquitous Computing, Electronics &amp; Mobile Communication Conference (UEMCON), 0059–0066. https://doi.org/10.1109/uemcon54665.2022.9965696

S. P. R. R. Raj and P. K. Gupta, "Vision Transformers in Medical Image Analysis: A Review," IEEE Transactions on Medical Imaging, vol. 41, no. 9, pp. 2161- 2175, Sep. 2022. https://www.semanticscholar.org/paper/Transformers-in-Medical-Image-Analysis:-A-Review-He-Gan/42bad1b72259aa1ff70d7ce2539220a83f1af9a4

Saraf, P., Tharaniesh, P. R., & Singh, S. (2024). Skin disease detection using convolutional neural network. 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), 1– 6. https://doi.org/10.1109/nmitcon62075.2024.10699196

Upadhyay, A., Singh, G., Mhatre, S., & Nadar, P. (2023). Dog skin diseases detection and identification using convolutional neural networks. SN Computer Science, 4(3). https://doi.org/10.1007/s42979-022-01645-5