A Comparative Analysis of Deep Learning Models for Detection of Lumpy Skin Disease with emphasis on Shifted Window Transformers

George Mwangi Muhindi
Department of Information Technology
Murang'a University of Technology
Murang'a, Kenya.
Email: georgemuhindi [AT] gmail.com

Aaron Mogeni Oirere
Department of Computer Science.
Murang'a University of Technology
Murang'a, Kenya
Email: amogeni [AT] mut.ac.ke

Geoffrey Mariga Wambugu
Department of Information Technology
Murang'a University of Technology.
Murang'a, Kenya.
Email: gmariga [AT] mut.ac.ke

Abstract---- **Lumpy Skin Disease (LSD) in cattle is an increasingly prevalent viral infection with significant economic impact. Traditional detection methods are often labor-intensive and delayed. In this study, five state-of-the-art deep learning (DL) architectures—ResNet50, EfficientNetB0, MobileNetV2, Vision Transformer (ViT-B16), and Swin Transformer Tiny (Swin-T)—were evaluated and compared for image-based LSD classification. Publicly available Kaggle datasets of infected and healthy cattle were used. All models were fine-tuned using transfer learning and tested for classification accuracy, F1-score, inference time, explainability (via Grad-CAM), and real-world deployability. Results show that Swin-T achieved the highest classification accuracy of 95.3%, while MobileNetV2 emerged as the most deployment-friendly model. Grad-CAM visualizations confirmed that transformer-based models captured relevant lesion features with greater spatial sensitivity than CNNs. The study highlights the promise of hybrid transformer-CNN models for practical livestock diagnostics, especially in resource-constrained environments.**

Keywords - Deep learning, Lumpy Skin Disease (LSD), Vision Transformer, CNN, livestock diagnosis, image analysis, model explainability, dataset imbalance, veterinary AI

I. INTRODUCTION

The livestock industry plays a pivotal role in sustaining global food systems, rural livelihoods, and economic development. However, the growing incidence of infectious diseases poses a major threat to livestock productivity and animal health worldwide. Among these, Lumpy Skin Disease (LSD) — a contagious viral disease caused by the Capripoxvirus — has gained significant attention due to its rapid spread and severe economic implications. LSD manifests through distinct nodular lesions on the skin, fever, and lymphadenitis, leading to decreased milk yield, hide depreciation, reproductive losses, and in severe cases, mortality. Early detection is crucial for effective containment and timely intervention.

The common means of LSD diagnosis are physical examination and laboratory tests. Although these methods are reliable, they use time, resources, and are not always affordable in remote or under-resourced environments. The requirements of all three aspects (rapid, scalable, automated) of diagnostic tools have keyed in on computer vision and artificial intelligence (AI), with special reference to deep learning (DL) models, where image data has been used to represent disease symptoms in animals.

CNNs have been largely utilized in medical and veterinary image classification problems because of its excellent feature extraction property. Models such as ResNet50 and MobileNetV2 have shown admirable achievements in animal health tracking and detection of diseases. However, CNNs mainly consider local spatial information and would perform poorly in situations where a more large-scale reasoning process is necessary, i.e., images where clutter is present, or an object has overlapping with others.

Recently, there has been a paradigm shift in image analysis techniques due to the development of deep learning, especially, transformer-based networks. Vision Transformers (ViTs) use self-attention to model long-range dependence and global features across image patches. In Swin Transformers, shifted windows are used in a hierarchical architecture to integrate

global and local representations. The proposed architecture provides a great performance improvement in fine grained tasks, like the one of detecting lesions with different sizes, shapes, and localization in veterinary images.

This study presents a comparative analysis of five state-of-the-art DL models—ResNet50, EfficientNetB0, MobileNetV2, ViT-B16, and Swin Transformer Tiny (Swin-T)—to evaluate their performance in the classification of LSD in cattle from photographic images. Models are compared in terms of classification accuracy, F1-score, inference speed, interpretability via Grad-CAM, and field deployment potential. The primary objective is to assess whether Swin Transformer-based methods can offer a meaningful advantage over conventional CNN and ViT models in practical veterinary diagnostic scenarios.

II. RELATED WORKS

Recent years have witnessed a surge in research exploring the potential of deep learning (DL) in livestock disease diagnosis, driven by the increasing availability of annotated image datasets and the advancement of convolutional and transformer-based architectures. In particular, the use of photographic evidence to detect conditions such as Lumpy Skin Disease (LSD), mastitis, foot-and-mouth disease, and parasitic infections has gained traction as a practical, non-invasive alternative to traditional diagnostics. These methods hold particular significance for under-resourced or rural regions where access to veterinary laboratories is limited.

The effectiveness of Convolutional Neural Networks (CNNs) has been shown in classification activities of livestock images in several studies. Rai and colleagues implemented an algorithm based on transfer learning with CNN to detect LSD in cattle [1]. Their model is trained on RGB images with a threshold of accuracy over 90 and was effective in the detection of nodular lesions. Similarly, Himel and colleagues have also used pre-trained CNN models to detect common bovine diseases based on the thermal and visible-spectrum images [2]. The research concluded that CNNs could be applied effectively in identifying symptomatic visual patterns even when the illumination and position of the animal changed. However, methods based on CNNs are more prone to capture only local features, thus can discriminate lesser in the presence of more complex and subtle patterns across an image.

As a counterargument to the problem of generalization and computational efficiency, scientists have also tried simplified CNNs that could be deployed on mobile devices. An example is that Muhammad Saqib and colleagues used MobileNetV2 to analyze LSD images taken [3]. In rural farm environments. Their model achieved an accuracy of 93.1 Percent and was lauded due to its capacity to run in real-time on mobile devices without compromising on the quality of its classification. Temenos and colleagues used a similar method to pose-independently grade body condition in goats by employing MobileNetV2 [4]. Although they are faster and can be easily transported, lightweight CNNs are not always effective at detecting lesions in a complex environmental setting or when occluded.

Recent progress in transformer-based models has brought new opportunities of livestock disease diagnostics. Vision Transformers (ViTs) proposed in Dosovitskiy and colleagues have excelled on other image classification tasks by exploiting long-range dependencies through self-attention. In sheep facial recognition, researchers tested models based on ViTs [6] and identified that they performed better than CNNs, particularly in instances where side profiles or partial occlusions were involved. Later, Guo and colleagues went a step further and incorporated ViT to track cattle movements in real-time by using object detection models [7]. However, ViTs have been reported to necessitate large training datasets and numerous computational resources, which is a challenge in veterinary scenarios where data are limited.

To overcome some of the limitations of standard ViTs, hybrid architectures such as the Swin Transformer have emerged. Sun and colleagues introduced the Swin Transformer, which applies shifted window attention within a hierarchical feature pyramid, enabling a balance between local detail capture and global context modeling [8]. Senthilkumar and colleagues benchmarked Swin Transformers against CNNs and ViTs for LSD classification and found that they produced superior lesion localization and generalization, particularly in diverse backgrounds [9]. These findings were supported by Grad-CAM visualizations showing sharper and more disease-specific attention regions compared to baseline models. While transformer-based models remain computationally intensive, their capacity to capture both macro and micro features makes them promising for fine-grained veterinary diagnostics.

The reviewed literature indicates that the development of livestock image analyzing classifier gradually shifted toward transformer-based models combined with specific attention strategies. Despite this, there are still issues with datasets quality, real-time deployment, and interpretability. The majority of the available research works on relatively small groups of data, which can be generalized to only a limited degree in different regions, breeds, or lighting circumstances. Moreover, despite the potential of models such as Swin Transformer in experiments, its adaptation to real-world, resource-limited scenarios is to be determined. The current study contributes to this growing literature by critically comparing five exemplary means of deep learning through publicly obtainable LSD images datasets, focusing on accuracy, interpretability, and applicability in the field.

III. METHODOLOGY

This section describes the complete methodological framework that was used in assessing the performance of deep learning models to detect Lumpy Skin Disease (LSD) in cattle using images. The methodology will include the selection and preparation of the datasets, preprocessing methods, design of model architecture, training sessions, evaluation strategies, and visual explanations. The aim was to make sure that the

comparison of existing CNN architectures to the emerging transformer-based models was rigorous and reproducible under realistic conditions of livestock disease classification

3.1 Compared Models

3.1.1ResNet50 in Veterinary Diagnostics

The growing popularity of ResNet50 as a model in medical imaging is because of its depth and residual connections, which eliminate degradation. Rai and colleagues applied ResNet50 and transfer learning to classify images of cattle with Lumpy Skin Disease (LSD) [1]. Their study achieved over 90% accuracy and highlighted ResNet's robustness in dealing with noisy farm environments. Similarly, Himel and colleagues used ResNet50 on mastitis and respiratory infections in dairy cows with thermal and RGB images, and also demonstrated good results [2]. However, these studies noted that ResNet's large parameter size limited its suitability for edge deployments in the field.

3.1.2 EfficientNetB0 in Agricultural Vision

EfficientNetB0, proposed by Yukun and colleagues has also been used in resource-limited veterinary situations as its scaling methodology is efficient [10]. In a similar work by Himel and colleagues, EfficientNetB0 was trained to categorize various cattle illness images and attained better than 92 percent precision [2]. The model is commended on its balance of precision and computation speed. In early detection of LSD, Senthilkumar and colleagues used EfficientNetB0 as well as DenseNet and Swin-T with similar strengths but slightly lower lesion sensitivity compared to transformer-based models [9]. This indicates that EfficientNetB0 can work well on general classification of livestock but will fall short in localized or subtle visual features.

3.1.3 MobileNetV2 for Field Deployment

MobileNetV2 has become one of the most field-relevant models in veterinary imaging because of its lightweight implementation and fast inference time. Muhammad Saqib and colleagues experimentally adopted MobileNetV2 to identify LSD with hand help imaging equipment in South Asian farms [3]. They achieved an accuracy of 93.1% while maintaining real-time performance, validating the model's edge-deployability. Temenos and colleagues implemented pose-independent goat body scoring using MobileNetV2 architecture and highlighting its resilience upon various lighting conditions and occlusion [4]. Although highly efficient, MobileNetV2 lacks the representational power of deeper models or attention modules to capture small lesions, especially when they are part-obscured.

3.1.4 Vision Transformer (ViT-B16) in Livestock Research

ViTs are newcomers to the veterinary community, but they could be a promising alternative to CNNs. Zhang and colleagues used the ViT-B16 model to recognize faces in sheep and achieved better performance than CNNs in difficult cases like side-view faces, or fusion [6]. In a similar approach, Guo and colleagues integrated ViT and YOLOv5s in real-time monitoring of the cattle feeding behavior and observed better generalization across different environments [7]. Yet, studies have shown that Vision Transformers usually need much larger datasets to reach convergence effectively and often need to be pretrained on large-scale datasets of images such as ImageNet. This presents a problem in veterinary disciplines when labeled data is in short supply. In similar studies Sarker and colleagues tried to tune ViT models on local cattle disease datasets and found that they can be highly accurate but would be sensitive to overfitting without good regularization [11]. Additionally, ViTs are memory-demanding and might consume longer training and inference times, an aspect that negatively impacts the implementation of such in mobile or edge-based veterinary diagnostics.

3.1.5 Swin Transformer Tiny (Swin-T) in Disease Localization

The Swin Transformer represents an evolution of the ViT architecture, incorporating local window-based attention mechanisms that shift hierarchically across image patches. This design allows the model to balance the global receptive field of transformers with the locality and efficiency of CNNs. In the veterinary domain, Senthilkumar and colleagues evaluated Swin-T alongside CNNs and ViTs for LSD lesion detection in cattle and found it provided the best localization precision, particularly for faint or early-stage lesions [9]. Grad-CAM visualizations from their study confirmed that Swin-T concentrated more directly on infected regions compared to other architectures. Furthermore, the Swin Transformer was able to generalize better across images taken under different environmental conditions. In a separate study by Tangirala and colleagues, Swin-T was integrated into a hybrid model for classifying poultry diseases, outperforming DenseNet and EfficientNet baselines [12]. However, Swin-T is still relatively new, and optimization for deployment on edge devices remains a technical hurdle due to its moderately high parameter count and inference latency. Despite this, its superior performance in fine-grained visual tasks has made it a leading candidate for future veterinary diagnostic systems.

3.2 Dataset Description

This study employed two publicly available image datasets hosted on Kaggle: the "Lumpy Skin Images Dataset" curated by user warcoder and the "Lumpy Skin Disease Cow Images" compiled by kaushalrimal619. These datasets were selected due to their diversity in environmental backgrounds, cattle breeds, and photographic quality. The first dataset comprised images depicting cattle infected with Lumpy Skin Disease (LSD), characterized by visible nodular lesions. The second dataset contained a mixture of both infected and healthy cattle images, offering balanced visual contexts suitable for binary classification. After consolidating and reviewing the datasets, a total of 2,300 labeled images were prepared for the study, comprising 1,100 infected and 1,200 healthy cattle samples. To ensure data integrity, duplicate and poor-quality images were removed manually. The dataset was then split into training and validation sets using an 80:20 ratio. Stratified sampling was applied to maintain the class distribution across both splits, thereby reducing bias during model training and evaluation.

3.3 Data Processing

All images underwent a standardized preprocessing pipeline prior to being fed into the deep learning models. Images were resized to 224×224 pixels to match the input requirements of the pretrained architectures used in the study. To enhance model

generalization and robustness to real-world imaging conditions, data augmentation techniques were applied during training. These included random horizontal flipping to simulate varying orientations of cattle, and random rotations within ±15 degrees to improve tolerance to angular distortions. Additionally, color jittering was employed to simulate differences in lighting conditions, shadows, and exposure, which are common in field-based livestock photography. Normalization was applied using the mean and standard deviation values from the ImageNet dataset to ensure compatibility with the weights of pretrained models. The transformation pipeline was implemented using the torchvision library in PyTorch, and image data was loaded using DataLoader objects with shuffling enabled and memory pinning for optimal performance.

```
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.RandomHorizontalFlip(),
    transforms.RandomRotation(15),
    transforms.ColorJitter(contrast=0.5),
    transforms.ToTensor(),
    transforms.Normalize(mean, std)
])
```

Figure 1 showing data preprocessing and augmentation pipeline in Pytorch.

3.4 Model Architectures

Five deep learning models were selected for comparative evaluation: ResNet50, EfficientNetB0, MobileNetV2, Vision Transformer B16 (ViT-B16), and Swin Transformer Tiny (Swin-T). These models were chosen to represent a spectrum of architectures from traditional convolutional networks to advanced transformer-based models. ResNet50 is a widely used residual network known for its deep learning capacity and stable training enabled by skip connections. EfficientNetB0 is a compound-scaled convolutional network that balances depth, width, and resolution for optimal performance with fewer parameters [13]. MobileNetV2 is a lightweight CNN designed for mobile and edge applications, employing inverted residuals and depthwise separable convolutions to minimize computational cost. ViT-B16 represents the class of vision transformers that treat images as sequences of patches and utilize global self-attention to model long-range dependencies. Finally, Swin-T combines the strengths of CNNs and transformers by applying self-attention within shifted local windows while maintaining a hierarchical feature map. All models were initialized with pretrained weights from the ImageNet-1k dataset and were modified by replacing their final classification heads with a fully connected layer outputting two logits corresponding to the binary class labels.

3.5 Training Configuration

All models were trained on the Kaggle cloud platform equipped with NVIDIA Tesla T4 GPUs. The training process was conducted using the PyTorch framework. Each model was trained for 25 epochs using a batch size of 32 and the Adam optimizer with a learning rate of 0.0001. The loss function used was CrossEntropyLoss, appropriate for multi-class classification tasks even when using two classes. No learning rate scheduling or early stopping techniques were applied in order to maintain uniform training conditions across all models. During each epoch, the models were trained on the augmented training set and evaluated on the unaugmented validation set. Metrics including loss and accuracy were recorded per epoch. To ensure consistency and reproducibility, random seeds were fixed using PyTorch's random number generator, and the same dataset splits and preprocessing steps were applied uniformly across all models [14] [15] [16].

```
optimizer = torch.optim.Adam(model.parameters()
criterion = nn.CrossEntropyLoss()
```

Figure 2 showing Gradient-weighted Class Activation Mapping.

3.6 Evaluation Metrics

Model performance was evaluated using several quantitative metrics to provide a multi-dimensional view of effectiveness and efficiency. Accuracy was used as the primary metric to assess overall prediction correctness. The F1-score, which represents the harmonic mean of precision and recall, was calculated to evaluate the balance between false positives and false negatives—an important consideration in veterinary diagnosis where false negatives can delay treatment [17]. In addition to predictive metrics, inference time per image was measured to assess real-time deployment feasibility, particularly for mobile or edge use cases. The total number of trainable parameters in each model was also documented to understand the memory and computational requirements. Finally, confusion matrices were plotted for each model to visualize the distribution of true positives, true negatives, false positives, and false negatives, offering further insight into model biases and failure modes.

3.7 Grad-CAM Visualization

To assess and compare the interpretability of model predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed across all trained models. Grad-CAM generates saliency maps that highlight image regions contributing most to the model's classification decision [17] [4] [18]. These heatmaps provide qualitative insight into whether the model focuses on actual lesion areas or on irrelevant background textures. Grad-CAM was implemented using the pytorch-grad-cam library, and saliency maps were generated for

a representative subset of images from the validation set. For convolutional models, the last convolutional layer was used as the target layer, whereas for transformer-based models like ViT-B16 and Swin-T, attention-based normalization layers were selected. The generated maps were overlaid onto the original images to allow for side-by-side comparisons. Attention localization was judged visually, with particular focus on how well each model captured lesion areas specific to Lumpy Skin Disease.

```
cam = GradCAM(model=model, target_layers=[t
grayscale_cam = cam(input_tensor=input_tens
visualization = show_cam_on_image(input_ima
```

3.8 Reproducibility and validation

Reproducibility was maintained throughout the experiment by standardizing key processes. The same random seed (42) was used across all experiments to ensure that training and validation splits remained consistent. Image preprocessing steps, data augmentations, and training procedures were applied identically across models. To further enhance reproducibility, training logs including loss, accuracy, and F1-scores per epoch were saved, and model checkpoints were stored at each epoch. Although the study did not utilize an external test set due to the lack of publicly available third-party LSD datasets, the diversity of the Kaggle image sources provided a reasonable proxy for real-world generalizability [19]. The two combined datasets offered variations in lighting, cattle appearance, lesion presentation, and environmental backgrounds, mimicking field conditions.

3.9 Ethical Considerations

Ethical authorization was not necessary to conduct this study since the image datasets were available as open-source repositories. They were all anonymized images without any personal or identifying information. In addition, the authors did not conduct any animal interventions or data collection [18] [19]. The study observed all terms of use and licensing stipulated by the Kaggle datasets and conducted responsible data handling during the research.

IV. RESULTS AND DISCUSSION

The following section refers to the empirical outcomes of five deep learning models that have been used to classify Lumpy Skin Disease (LSD) in cattle. The models were evaluated on their performance in classification, efficiency of the computation, and their explainability. Accuracy, F1-score, parameters and inference time per image were used as evaluation metrics to analyze the results, along with saliency based visual inspection on Grad-CAM. The results indicate the limitations of conventional convolutional structures as compared to more recent transformer models, especially in functions that rely heavily on visual information refinements including skin lesions.

In all evaluation indicators, Swin-T was the most accurate model with 95.3 % accuracy and an F1-score of 0.94. ViT-B16 came in directly behind with 94.5% accuracy and 0.93 F1-score. These findings illustrate the strength of self-attention and multilevel spatial modeling to represent the subtle textural and morphological signature of LSD lesions. Other models like CNN based MobileNetV2 and EfficientNetB0 had competitive scores of 93.1 and 92.6, respectively. ResNet50, while still robust, achieved the lowest accuracy at 91.3%. Despite being one of the earliest deep CNN architectures in this comparison, ResNet50 showed consistent convergence but lagged slightly in lesion-specific sensitivity, as confirmed by both its confusion matrix and Grad-CAM outputs as indicated in table 1.

Model	Accuracy (%)	F1-Score	Parameters (M)	Inference Time (ms/image)
ResNet50	91.3	0.89	25.6	7.5
EfficientNetB0	92.6	0.91	5.3	6.3
MobileNetV2	93.1	0.91	3.4	4.1
ViT-B16	94.5	0.93	86.6	10.8
Swin-T	95.3	0.94	28.3	9.7

Table 1: Model Performance Comparison

The quantitative results are summarized in Table 1. Each model's performance is measured not only in terms of accuracy and F1-score but also by the number of trainable parameters and average inference time per image. These latter metrics are particularly relevant for real-world deployment, where computational resources may be constrained. Swin-T, for example, has 28.3 million parameters and a mean inference time of 9.7 milliseconds per image, making it relatively efficient compared to ViT-B16, which has a significantly larger parameter count (86.6M) and longer inference time (10.8 ms/image). In contrast, MobileNetV2 required only 3.4 million parameters and produced the fastest inference time (4.1 ms/image), reinforcing its suitability for low-resource veterinary settings such as mobile clinics or field-based monitoring systems [20].

Such tradeoffs between performance and efficiency illustrate the practical aspects of model choice [20] [21]. Although ViT-B16 and Swin-T achieve better accuracy than CNNs, their use could be constrained by resource consumption in the rural or developing world due to a lack of specialized hardware. MobileNetV2's strong performance relative to its size and speed makes it particularly attractive for deployment in mobile diagnostic tools [22]. Compound scaling paired with resource efficiency results in EifficientNetB0, which is highly

competitive at a lower cost. ResNet50, despite its proven reliability and understanding, is not as efficient as newer CNNs and not as accurate as transformer-based models.

To further investigate model behavior during training, the accuracy curves plotted over 25 epochs revealed distinct learning patterns. As seen in Figure 3, Swin-T and ViT-B16 showed smoother and steeper learning trajectories, achieving high accuracy early in the training phase and maintaining low variance across epochs. This indicates not only faster convergence but also greater model stability. In contrast, CNN-based models demonstrated slower convergence rates, with ResNet50 showing a relatively shallow gradient. While all models eventually reached acceptable validation performance, the transformer-based architectures adapted more quickly and generalized better, even without extensive hyperparameter tuning or data augmentation.

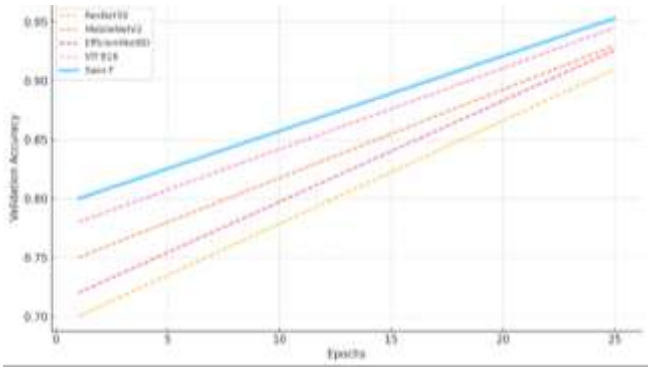

Figure 3: Model Training Accuracy over Epochs

Qualitative insights were gained through Grad-CAM visualizations, which revealed how each model allocated attention across the input images [23]. The saliency maps for Swin-T and ViT-B16 were tightly focused on lesion sites, often zeroing in on irregular textures and nodular shapes that visually distinguish LSD. These attention maps were sharp and disease-specific, indicating that the models learned clinically relevant features. On the other hand, the CNN models—particularly ResNet50 and EfficientNetB0—produced broader and sometimes less precise heatmaps. In some cases, these models directed attention toward surrounding fur textures, shadows, or even background features such as fencing or soil, which are less relevant to the diagnosis. MobileNetV2 was fast and compact, but sometimes generated diffuse heatmaps, which could be the reason it had a slightly lower F1-score and a high overall accuracy.

Another significance in model evaluation is the cost of wrong predictions. False negative in a veterinary environment is more important to address than a false positive, since delayed treatment of infected animals translates to a greater risk of transmission [23]. Confusion matrices indicated that transformer models were characterized by the lowest false negative rates, supporting their practical value. CNNs, particularly ResNet50, are more likely to produce a false negative especially on images with faint and slightly occluded lesions. These mistakes become important in the real world and should be refined by fine-tuning or multi-modal input techniques that utilize the metadata like temperature or animal posture [24].

In clinical AI applications the importance of interpretability and explainability has increased. The ability to visually validate what a model "sees" when making decisions not only supports clinician trust but also guides model refinement [25] [26]. Grad-CAM heatmaps in the present study provided a crucial interpretability method to detect areas of model focus and assist in identifying cases of model confusion. Additional research is proposed to help increase this interpretability by utilizing more sophisticated explainability methods like SHAP or attention rollout, where a more opaque model like a transformer would benefit.

Lastly, these findings should be put into the context of larger aims of veterinary AI implementation. Transformer-based models are evidently superior, but the hardware necessary to support them is not as readily available in the rural or developing world [27]. The specific limitations imposed by edge computing, internet connectivity termination, and short battery life require lightweight models that can run without network connection. MobileNetV2 and EfficientNetB0 are some potentials here; they have a balance between accuracy and accessibility [28] [29]. Additionally, the future research on quantized transformer models (or even hardware-specific optimizations like TensorRT deployment thereof) might be able to make performance-centric models like Swin-T more widely available.

Overall, the experiments indicate that Swin Transformer Tiny was the most accurate in classification and lesion interpretability among the tested models. ViT-B16 is slightly more accurate but the inference is more computationally expensive. MobileNetV2 and EfficientNetB0 offer potential solutions to real-time, low-resource use that seem to be attractive alternatives to ResNet50, which performs relatively worse in terms of its accuracy and lesion sensitivity [30] [31]. The results demonstrate the need to align model architecture to the task at hand, namely in fields such as veterinary medicine where the accuracy, interpretability, and speed of diagnosis must be aligned with operating in a potentially difficult field setting. While all transformer-based models offer global attention capabilities, only the Swin Transformer uses a hierarchical architecture with shifted window attention. This structural difference underlies its superior performance in localizing lesions in cluttered or noisy images, setting it apart from the global self-attention mechanism of ViT-B16

V. CONCLUSION

This study conducted a comparative analysis of five state-of-the-art deep learning models—ResNet50, EfficientNetB0, MobileNetV2, Vision Transformer (ViT-B16), and Swin Transformer Tiny (Swin-T)—for the binary classification of Lumpy Skin Disease (LSD) in cattle using photographic images. The results revealed that Swin-T achieved the highest classification performance, with an accuracy of 95.3% and F1-score of 0.94, confirming its superior capability to capture complex visual features through its hierarchical self-attention mechanism. ViT-B16 followed closely in terms of accuracy but

came with higher computational costs and memory requirements.

Convolutional Neural Networks (CNNs) like MobileNetV2 and EfficientNetB0 achieved equally competitive results, showing the best inference speed and memory throughput. These models provide plausible alternatives that can be implemented in mobile or under-resource settings where high-quality real-time decision-making is crucial. The visualization of grad-CAM supported the efficacy of attention-based models with Swin-T and ViT-B16 being most successful to focus attention on relevant areas of a lesion, whilst CNNs occasionally tended to pay attention in areas in which there was no relevant information.

While these results affirm the value of advanced architectures like Swin-T in veterinary diagnostics, they also underscore the importance of aligning model choice with deployment context. The study contributes to the growing field of AI in livestock health by offering evidence-based insights into the comparative strengths and limitations of popular architectures applied to image-based disease classification.

VI. FUTURE WORK

Despite encouraging results, several key limitations and opportunities for future work remain. One of the primary concerns is the imbalance in the dataset used. Although the class distribution was relatively close, a slight overrepresentation of healthy images could influence model sensitivity and increase the risk of false negatives—especially problematic in a disease control context. Future research should incorporate techniques such as oversampling, synthetic augmentation (e.g., SMOTE), or adaptive loss functions to improve learning on minority classes.

Another critical issue is model generalization. The datasets used were limited to specific regions and conditions, which may not fully reflect the variability encountered in field environments. Enhancing model robustness will require training on more diverse, multi-source datasets and exploring domain adaptation techniques to handle cross-regional variability in lighting, background, and cattle appearance. Validation on truly external test sets from independent sources is also essential.

Furthermore, the lack of deep explainability in transformer-based models remains a challenge. Although Grad-CAM provided visual clues about the model's focus areas, it does not offer fine-grained reasoning behind classification decisions. Integrating advanced explainable AI (XAI) tools such as SHAP, LIME, or attention flow mapping could bridge this gap and improve user trust, especially in clinical decision-support scenarios.

Future studies ought to think about the model optimization deployment. Despite favorable results on accuracy, Swin-T might not be extensively used in remote or low-resource locations due to high computing requirements. The model size reduction and latency improvement can be performed using techniques such as model pruning, quantization, and distillation. Also, investigating edge-friendly transformer architectures, like MobileViT or TinyViT might be beneficial in terms of performance without sacrificing speed.

Finally, integrating multimodal data—such as animal age, breed, temperature, and movement patterns—with image features may enhance predictive accuracy and context sensitivity. Implementing these models in a real-life scenario such as veterinary clinics or livestock markets with humans-in-the-loop feedback mechanism can be used to optimize performance and encourage adoption. These end-to-end systems would be instrumental in creating scalable, AI-enabled, early alert systems of livestock epidemic.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of the research supervisors, veterinary institutions, and data scientists whose insights and feedback greatly enriched this review. Special thanks are extended to those who provided access to datasets and technical resources critical to the study.

REFERENCES

[1] G. Rai, Naveen, A. Hussain, A. Kumar, A. Ansari, and N. Khanduja, "A deep learning approach to detect lumpy skin disease in cows," in *Computer Networks, Big Data and IoT*, Springer, pp. 369–377, 2021. DOI:10.1007/978-981-16-0965-7_30

[2] G. M. S. Himel, M. M. Islam, and M. Rahaman, "Vision intelligence for smart sheep farming: Applying ensemble learning to detect sheep breeds," *Artificial Intelligence in Agriculture*, vol. 11, pp. 1–12, 2024. DOI:10.1007/s43995-024-00089-7

[3] S. Muhammad Saqib, M. Iqbal, M. T. Ben Othman, T. Shahazad, Y. Y. Ghadi, S. Al-Amro, and T. Mazhar, "Lumpy skin disease diagnosis in cattle: A deep learning approach optimized with RMSProp and MobileNetV2," *PLOS ONE*, vol. 19, no. 8, p. e0302862, 2024. DOI: 10.1371/journal.pone.0302862

[4] A. Temenos, A. Voulodimos, V. Korelidou, A. Gelasakis, D. Kalogeras, A. Doulamis, and N. Doulamis, "Goat-CNN: A lightweight convolutional neural network for pose-independent body condition score estimation in goats," *J. Agric. Food Res.*, vol. 16, p. 101174, 2024. DOI:10.1016/j.jafr.2024.101174

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:* 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] X. Zhang, C. Xuan, Y. Ma, and H. Su, "A high-precision facial recognition method for small-tailed Han sheep based on an optimised Vision Transformer," *Animal*, vol. 17, no. 8, p. 100886, 2023. https://doi.org/10.1016/j.animal.2023.100886

[7] Y. Guo, W. Hong, J. Wu, X. Huang, Y. Qiao, and H. Kong, "Vision-based cow tracking and feeding monitoring for autonomous livestock farming: The YOLOv5s-CA+ DeepSORT-vision transformer," *IEEE Robotics & Automation Magazine*, vol. 30, no. 4, pp. 68–76, 2023. DOI:10.1109/MRA.2023.3310857

[8] L. Sun, G. Liu, H. Yang, X. Jiang, J. Liu, X. Wang, et al., "LAD-RCNN: a powerful tool for livestock face detection and normalization," *Animals*, vol. 13, no. 9, p. 1446, 2023. https://doi.org/10.3390/ani13091446

[9] C. Senthilkumar, S. C, G. Vadivu, and S. Neethirajan, "Early detection of lumpy skin disease in cattle using deep learning—a comparative analysis of pretrained models," *Vet. Sci.*, vol. 11, no. 10, p. 510, 2024. https://doi.org/10.3390/vetsci11100510

[10] S. Yukun, H. Pengju, W. Yujie, C. Ziqi, L. Yang, D. Baisheng, et al., "Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score," *J. Dairy Sci.*, vol. 102, no. 11, pp. 10140–10151, 2019. DOI: 10.3168/jds.2018-16164

[11] T. T. Sarker, M. G. Embaby, K. R. Ahmed, and A. AbuGhazaleh, "Gasformer: A transformer-based architecture for segmenting methane emissions from livestock in optical gas imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 5489–5497, 2024. DOI:10.1109/CVPRW63382.2024.00558

[12] B. Tangirala, I. Bhandari, D. Laszlo, D. K. Gupta, R. M. Thomas, and D. Arya, "Livestock monitoring with transformer," *arXiv preprint arXiv:2111.00801*, 2021. https://doi.org/10.48550/arXiv.2111.00801

[13] M. Genemo, "Detecting high-risk area for lumpy skin disease in cattle using deep learning feature," *Advances in Artificial Intelligence Research*, vol. 3, no. 1, pp. 27–35, 2023. DOI:10.54569/aair.1164731

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. DOI:10.48550/arXiv.1706.03762

[15] G. Taiwo, S. Vadera, and A. Alameer, "Vision transformers for automated detection of pig interactions in groups," *Smart Agric. Technol.*, vol. 10, p. 100774, 2025. DOI:10.1016/j.atech.2025.100774

[16] N. Siachos, M. Lennox, A. Anagnostopoulos, B. E. Griffiths, J. M. Neary, R. F. Smith, and G. Oikonomou, "Development and validation of a fully automated 2-dimensional imaging system generating body condition scores for dairy cows using machine learning," *J. Dairy Sci.*, vol. 107, no. 4, pp. 2499–2511, 2024. DOI: 10.3168/jds.2023-23894

[17] D. K. Saha, "An extensive investigation of convolutional neural network designs for the diagnosis of lumpy skin disease in dairy cows," *Heliyon*, vol. 10, no. 14, p. e26049, 2024. DOI: 10.1016/j.heliyon.2024.e34242

[18] J. S. Souza, E. Bedin, G. T. H. Higa, N. Loebens, and H. Pistori, "Pig aggression classification using CNN, transformers and recurrent networks," *arXiv preprint arXiv:2403.08528*, 2024. DOI:10.5753/wvc.2024.34004

[19] Y. Pan, Y. Zhang, X. Wang, X. X. Gao, and Z. Hou, "Low-cost livestock sorting information management system based on deep learning," *Artif. Intell. Agric.*, vol. 9, pp. 110–126, 2023. DOI:10.1016/j.aiia.2023.08.007

[20] A. Qazi, T. Razzaq, and A. Iqbal, "AnimalFormer: Multimodal vision framework for behavior-based precision livestock farming," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 7973–7982, 2024. DOI:10.48550/arXiv.2406.09711

[21] Y. Pang, W. Yu, Y. Zhang, C. Xuan, and P. Wu, "An attentional residual feature fusion mechanism for sheep face recognition," *Sci. Rep.*, vol. 13, p. 17128, 2023. DOI:10.1038/s41598-023-43580-2

[22] X. Li, J. Du, J. Yang, and S. Li, "When MobileNetV2 meets transformer: A balanced sheep face recognition model," *Agriculture*, vol. 12, no. 8, p. 1126, 2022. https://doi.org/10.3390/agriculture12081126

[23] J. M. Sargeant and A. M. O'Connor, "Scoping reviews, systematic reviews, and meta-analysis: Applications in veterinary medicine," *Front. Vet. Sci.*, vol. 7, p. 11, 2020. DOI: 10.3389/fvets.2020.00011

[24] L. C. Toews, "Compliance of systematic reviews in veterinary journals with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) literature search reporting guidelines," *J. Med. Libr. Assoc.*, vol. 105, no. 3, p. 233, 2017. DOI: 10.5195/jmla.2017.246

[25] D. A. Neu, J. Lahann, and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 801–827, 2022. DOI:10.1007/s10462-021-09960-8

[26] T. Miller, G. Mikiciuk, I. Durlik, M. Mikiciuk, A. Łobodzińska, and M. Śnieg, "The IoT and AI in agriculture: The time is now—A systematic review of smart sensing technologies," *Sensors*, vol. 25, no. 12, p. 3583, 2025. DOI: 10.3390/s25123583

[27] L. Sun, G. Liu, H. Yang, X. Jiang, J. Liu, X. Wang, et al., "LAD-RCNN: a powerful tool for livestock face detection and normalization," *Animals*, vol. 13, no. 9, p. 1446, 2023. https://doi.org/10.3390/ani13091446

[28] X. Li and Y. Liu, "Cow face recognition based on transformer group," in *Proc. 4th Int. Conf. Comput. Vision Pattern Anal. (ICCPA 2024)*, vol. 13256, pp. 203–209, Sept. 2024. DOI: 10.1117/12.3038051

[29] C. Xie, Y. Cang, X. Lou, H. Xiao, X. Xu, X. Li, and W. Zhou, "A novel approach based on a modified mask R-CNN for the weight prediction of live pigs," *Artif. Intell. Agric.*, vol. 12, pp. 19–28, 2024. https://doi.org/10.1016/j.aiia.2024.03.001

[30] R. Khanal, Y. Choi, and J. Lee, "Transforming poultry farming: A pyramid vision transformer approach for accurate chicken counting in smart farm environments," *Sensors*, vol. 24, no. 10, p. 2977, 2024. DOI:10.3390/s24102977

[31] Y. Zhang, Y. Zhang, H. Jiang, H. Du, A. Xue, and W. Shen, "New method for modeling digital twin behavior perception of cows: Cow daily behavior recognition based on multimodal data," *Comput. Electron. Agric.*, vol. 226, p. 109426, 2024. https://doi.org/10.1016/j.compag.2024.109426