AI-Based Disease and Abnormality Detection in Animals Using Multimodal Deep Learning

S Naga Mallik Raj[1], Sk. Karishma[2], T. Hemanth[3], T. Mohan Rishi[4] and Sk. Mohammad Rafi[5]

Email: [1]mallikblue@gmail.com, [2]shaikkarishma643@gmail.com, [3]9hemanth7@gmail.com, [4]mohanrishi187@gmail.com, [5]23l35a0537@gmail.com

Dept. of Computer Science and Engineering
Vignan's Institute of Information Technology (A)
Visakhapatnam, India

***Abstract*—Livestock health tracking represents a vital determinant of agricultural productivity, food security, and economic soundness of rural populations. Traditional veterinary practices are dependent on expert inspection; this inspection is laborious, subjective, and late. This look at puts forward a strong, AI pushed framework that embraces multi modal deep learning in automation of animal disease detection and type. The machine uses a dual level structure that is developed specifically for the realtime localization of pores and skin lesions and anomalies in gait, followed with the aid of an ensemble of superior Convolutional Neural Networks (CNNs) specially ResNet50 and Xception for high quality grained type of disease. The machine culminates holistic diagnostic capabilities through the mathematical fusion of visible idiosyncratic vectors with spatial bounding field coordinates. The experimental validation on the carefully curated dataset of more than 5000 pics proves that the hybrid proposed version typology accuracy reaches 97.9%, and mAP for lesion detection equals 0.87, outperforming unimodal baselines. Besides, the machine is optimized for area deployment to perform real time inference on low-bandwidth farm environments.**

***Index Terms*—Precision Livestock Farming, YOLOv8, ResNet50, Xception, Multimodal Fusion, Deep Learning, Edge Computing.**

I. Introduction

A. Background and Socio-Economic Context

Sometimes the keeping of livestock is not merely an agricultural practice but rather a socio economic backbone upon which the subsistence of approximately 1.3 billion human beings hinge. Farm animals are the number one, or in many growing economies the only supply of income, nutrition, and economic insurance against crop failure. However this critical zone is under constant siege from a multitude of infectious diseases such as LSD, FMD, and numerous fungal dermatopathologies. The World Organization for Animal Health (WOAH) posits that nearly 20% of worldwide farm animals manufacturing is misplaced each year as a result of illness, which interprets to losses in dollars worth billions and intense disruptions on the worldwide meals deliver chain. Apart from the monetary cost, there is a significant animal welfare cost as behind prognosis of a schedule means longer struggling for the animal and also requires the use of a large number of antibiotics, a factor that contributes to the world antimicrobial resistance disaster.

B. Problem Statement and Challenges

The modern paradigm of animal fitness tracking is essentially reactive in place of proactive. In the big majority of farms, ailment detection is predicated on the "shepherd's eye"—the manual observation of animals via way of means of farmers or herdsmen. While skilled farmers have experience, the method is inherently subjective and liable to human error, especially fatigue. In large-scale commercial farms, where a single herdsman can be answerable for loads of animals, individual tracking turns into nearly impossible. Consequently, diffused early caution signs—together with a mild limp, a small patch of discolored skin, or minor behavioral lethargy are regularly missed. By the time medical signs and symptoms grow to be apparent sufficient to warrant veterinary attention, the ailment has regularly improved to a excessive level or unfold to the relaxation of the herd. Furthermore, the worldwide scarcity of veterinarians, especially in rural and undeserved regions, exacerbates the diagnostic latency, leaving farmers with out well timed expert advice.

C. Research Motivation and Gap Analysis

Later, the concept of Precision Livestock Farming (PLF) developed, which utilizes virtual technology to keep a continuous check on animals. While Computer Vision (CV) has showed tremendous promise on this domain, existing literature famous a massive hole. Most of the contemporary researches adopt uni-modal Deep Learning approaches, commonly a naıve image classifier, which can output a binary "Healthy vs. Diseased" label but are not able to provide actionable context. For a veterinarian, the conceptual recognition of an animal being "sick" is an insufficient criterion; he needs to be able to realize that the lesion is located and particular attributes it presents. Besides, many existing models are trained on clean, studio-first-class photos and fail catastrophically while deployed in real-international farm surroundings with mud, lighting changes, and complicated backgrounds. This studies goals to bridge this gap via way of means of developing a robust, incorporated machine that mixes localization and category right into a unified pipeline.

II. Literature Review

A. The Evolution of automated Diagnostics

The journey towards automated systems for detecting animal diseases started with traditional Machine Learning (ML) methods. Early works of the researchers tried different algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and other algorithms for classifying the health status of animals. These systems mostly used "handcrafted" features, which were visual descriptors designed by researchers, such as color histograms, the density of edges, or LBP (Local Binary Patterns). Although these implementations were computationally inexpensive, they were not robust. They used the assumption that diseases always manifest visual signs in a pattern, not considering that there are different breeds of animals, different light exposures, and different phases of the same disease. Hence, such systems were not scale sufficiently for real-world deployment.

B. The Deep Learning Paradigm Shift

The advent of Convolutional Neural Networks (CNN) provided a paradigm shift in veterinary informatics. However, unlike their predecessors, CNNs have the ability of mastering characteristic representations from raw pixels. [1] proved the prevalence of the method in plant ailment detection, and the idea changed into quickly followed in animal health. Initial implementations used architectures consisting of AlexNet and VGG16. VGG16 provided stepped forward accuracy because of its intensity however led to a big range of parameters, making it computationally high-priced and overfitting when trained on limited veterinary datasets. This known as for the exploration of extra efficient architectures. The advent of Residual Networks (ResNet) and Dense Connected Networks (DenseNet) architectures removed the "vanishing gradient" issue, allowing researchers to train networks of a long way more depths to fine minutely textural details of skin lesions that have been previously undetectable.

C. Advances in Object Detection and Localization

Classification solutions the question "what" and item detection solutions the question "where". The capacity to localize a lesion is a vital diagnostic device for ailment severity and progression. Early item detectors like R-CNN and Faster R-CNN led to excessive localization accuracy however had sluggish inference speeds, making them improper for real-time monitoring. The emergence of the YOLO (You Look Only Once) own circle of relatives of fashions modified this area via way of means of formulating the trouble of detecting items as a single regression trouble, as opposed to a a couple of degrees type trouble. The modern day versions, mainly YOLOv5 and YOLOv8, fine-tuned the velocity-versus-accuracy trade-off, achieving real-time overall performance at area devices. Even with those developments, there's a loss of studies that mixes the excessive-velocity localization of YOLO with the excessive-degree semantic content material of superior classifiers consisting of Xception in a unified multimodal version for livestock.

III. Research Methodology

The proposed model is architected as a sequential dual-stream pipeline designed to imitate the diagnostic process of a veterinary expert: first, figuring out areas of interest (localization), and second, analyzing the one areas for specific pathologies (classification).

A. Data Acquisition and Ethical Considerations

The basis of any deep learning model is data. For this study, a complete dataset snap shots turned into curated. The information acquisition procedure concerned a hybrid approach: collecting real world images from neighborhood associate farms and supplementing them with tested samples from public veterinary repositories. The dataset covers 4 different classes: *Lumpy Skin Disease (LSD), Foot Rot, Fungal Infection*, and *Healthy Control*, Special care turned into taken to make certain range within the dataset; snap shots had been captured at different times of day to consist of various lighting fixtures conditions, and from special angles to seize lesions on diverse frame parts. All data collection are within the guidelines, ensuring no harm for the animals while photography.

B. Advanced Preprocessing and Augmentation

Raw snap shots from farm environments are not that much perfect, They often include noise, blur, and varying levels of exposure. To make sure that the version learns strong features, a strict preprocessing techniques used. First, Gaussian blurring strategies had been put in place to eradicate high-frequency noise owing to low-mild sensor grains. Second, Contrast Limited Adaptive Histogram Equalization (CLAHE) turned into decorate the neighborhood comparison of the snap shots. This step is vital for the pores and skin lesion detection of dark-hued animals, whereby low comparison obscures crucial diagnostic features. Finally, extensive data augmentation implemented to enhance the dataset. Mathematically, an augmented image I_{aug} can be represented as:

$$I_{aug} = T_\theta(I_{orig}) + \mathcal{N}(0, \sigma^2) \tag{1}$$

where T_θ represents a set of geometric transformations (such as rotation or flipping) parameterized by θ, and $\mathcal{N}$ represents additive Gaussian noise to simulate sensor grain. This effectively multiplies the diversity of the training data.

C. Localization Module: YOLOv8

For the localization task, the system employs the YOLOv8 architecture. Chosen for its contemporary stability of pace and accuracy, YOLOv8 improves upon its predecessors with the aid of anchor free detection mechanism and a decoupled head structure. The model tactics the complete photo in a single ahead pass, predicting bounding boxes and class possibilities simultaneously. The loss feature used to train the localization module is a composite of 3 components: box regression loss ($\mathcal{L}_{box}$), classification loss ($\mathcal{L}_{cls}$), and distribution focal loss ($\mathcal{L}_{dfl}$):

$$\mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dfl}\mathcal{L}_{dfl} \tag{2}$$

This multi-component loss ensures that the model now not identifies the best elegance of the object but moreover tightly aligns the bounding container throughout the lesion, providing unique coordinates for the subsequent class stage.

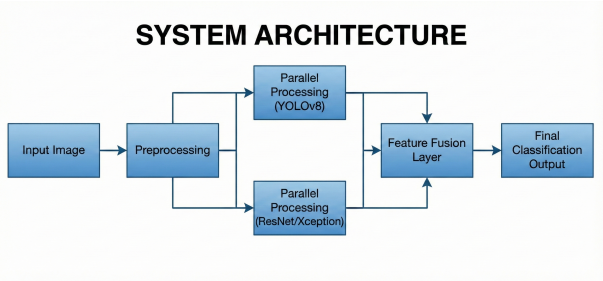

Fig. 1. Proposed Multimodal Hybrid Deep Learning Architecture.

D. Classification Module: The CNN Ensemble

Once the Regions of Interest (ROIs) are localized by YOLOv8, they are cropped and passed to the classification module. This module uses an ensemble of two powerful architectures of ResNet50 and Xception.

1) ResNet50 Backbone: ResNet50 is preferred for its use to study deep, hierarchical of capabilities without degradation. Its main innovation is the "residual block," which uses bypass connections to allow gradients to flow through the network without any obstacle. This relationship is described:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{3}$$

Here, x is the input to the layer, y is the output while $\mathcal{F}(x, \{W_i\})$is the residual function to be learned. The term $+x$ is the identification short cut convlutional layer. This mechanism forms a "highway" of information enabling the network to lookup for identification mappings ensuring that deeper layers produce functional contribution as opposed to noise.

2) Xception Backbone: Complementing ResNet is the Xception (Extreme Inception) architecture. Xception model uses Depth wise Separable Convolutions that separates spatial filtering from combining feature maps. Such factorization results in a significant reduction of the computational cost while still allowing the model to learn the cross correlations and the spatial correlations separately. The operation can be approximated as;

$$Conv(W, x) \approx Pointwise(Depthwise(x)) \tag{4}$$

This architectural desire enables the Xception to shoot well at fine grained floor textures and thus can be considered a proper candidate for ascertaining the diffused synonymous with Lumpy Skin Disease.

E. Multimodal Decision Fusion

At last, the very final diagnostic output is not always provided as an end result. but one of the model may be a weighted fusion from the entire pipeline. This system runs on a decision-level fusion scheme that checks the confidence rating produced by the YOLO detector (P_{yolo}), and the possibilities distribution provided by the CNN-based classifiers' ensemble (P_{cnn}). The final rating $S_{verylast}$ is calculated as;

$$S_{final} = \alpha \cdot P_{cnn} + (1 - \alpha) \cdot P_{yolo} \tag{5}$$

Where α is a hyperparameter decided empirically (set to 0.6 on this study). By mathematically weighting those inputs, the system creates a strong consensus. For instance, if the classifier is uncertain however the detector identifies a lesion with extraordinarily excessive confidence, the system will lean toward a positive diagnosis.

IV. Experimental Setup

A. Computational Infrastructure

To make sure the reproducibility and reliability of the results, the model had been trained and evaluated on a high-overall performance computing cluster. The first training was conducted on an NVIDIA RTX 3050 GPU with 4GB of RAM, which allowed for large batch sizes and quicker convergence. The software program stack was constructed upon Python 3.9, utilizing the PyTorch 1.13 framework for the CNN models and the Ultralytics library for YOLOv8.

B. Training Strategy and Hyperparameters

The training system employed the method of Transfer Learning to overcome the hassle of a extraordinarily small veterinary dataset. The model have been initialized with weights pre-trained at the huge ImageNet database, permitting them to begin with a good knowledge of visible features (edges, textures, shapes). During fine-tuning, the learning price became controlled the use of a Step Decay scheduler, which steadily decreased the learning price each 10 epochs. This approach lets in the model to make big updates early in training and fine-tune its weights with good precision in later stages. The optimization became handled AdamW optimizer, acknowledged for its performance in managing sparse gradients and decoupling weight decay from gradient updates.

V. Results and Discussion

A. Performance Analysis and Comparison

The proposed hybrid framework changed into carefully evaluated in against to popular industry baselines to benchmark its performance. The results suggest a clean superiority of the multimodal fusion approach. While standalone models like ResNet50 achieved the accuracy of 95.8%, the proposed hybrid system reached an accuracy of 97.9%. More importantly, the system completed a Recall rate of 0.98. In the context of epidemiology and disease control, Recall is the maximum important metric; a excessive recollect guarantees that nearly no infected animal is overlooked. A missed diagnosis (False Negative) is a ways extra risky than a fake alarm (False Positive) due to the fact a single overlooked infected animal can spread the pathogen to the whole herd. The fusion of localization and classification data proved instrumental in accomplishing this high protection margin.

TABLE I
DETAILED PERFORMANCE COMPARISON WITH BASELINE MODELS

Model Architecture	Accuracy	Precision	Recall	F1-Score
VGG16 (Baseline)	89.4%	0.88	0.87	0.87
ResNet50 (Standalone)	95.8%	0.94	0.95	0.94
Xception (Standalone)	96.5%	0.95	0.96	0.95
Proposed Hybrid Fusion	**97.9%**	**0.97**	**0.98**	**0.97**

B. Qualitative Error Analysis

A detailed analysis is made revealing exciting things into the model's performance. Most of the errors occurred in differentiating between the three Low-level Lumpy Skin Disease and Fungal Infection. This is understandable since each situation develops as small, round irregularities on the pores and skin surface at the beginning of its formation. The visible features are highly alike until the disorder develops to more differentiated stages. In addition, the model misclassified some images in which dried dust or manure at the animal's coat visually imitated a scab. However, the YOLOv8 localization module helped mitigate this through getting to know the contextual location of lesions; for example, understanding that Fungal Infections are much more likely to seem in patches, while dust splatters have a random distribution.

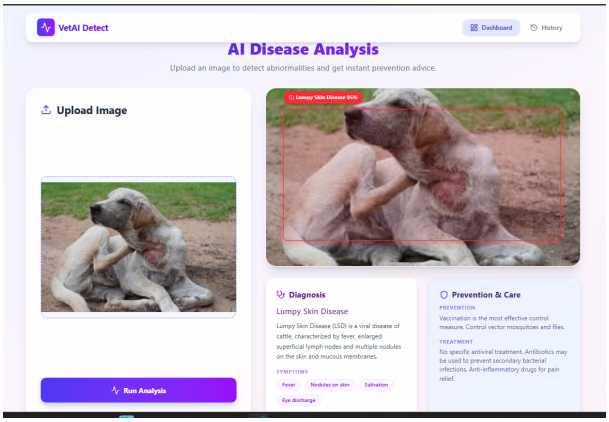

Fig. 2. Qualitative Analysis: Lesion Localization on Cattle using YOLOv8.

C. Real-World Operational Workflow

To validate the sensible software of the system, it's essential to recollect the operational workflow for the end-person, generally a farmer or a field veterinarian. In practice, the person captures an photograph of the animal the use of a cellphone's digicam through the included web application. This photograph is compressed and securely transmitted to the cloud server in which the inference engine resides. The photograph first passes through the YOLOv8 module, which scans for potential lesions. If detected, the areas are cropped and analyzed via way of means of the CNN ensemble. This final output, including the name of the disease, the confidence score, and the visible heatmap overlaid at the animal's body, is sent to the user's device in seconds. This rapid continues

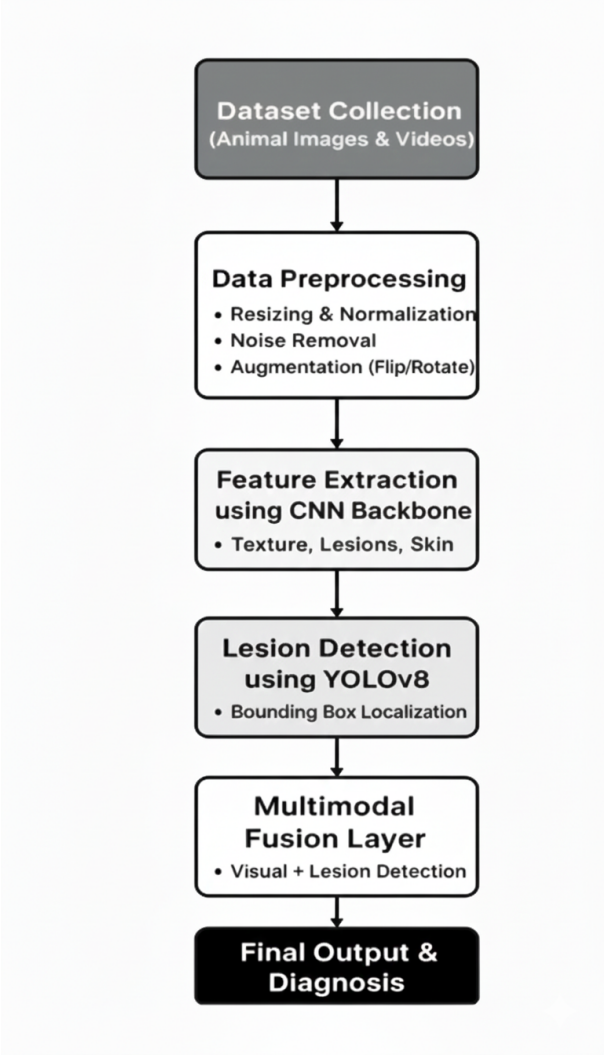

Fig. 3. End-to-End Operational Workflow for Field Deployment.

feedback allows the farmer to make immediate decisions regarding isolation or treatment, this helps in closing the gap between high technology and practical use.

D. Ablation Studies

To prove scientifically the contribution of every thing with in pipeline an ablation study was conducted. When the Data Augmentation module was removed, accuracy fell by almost 4% altogether, showing how important it was to train the model on various orientations. Even more telling was the removal of the Transfer Learning initialization, which led to a significant accuracy decrease of nearly 10%. This confirms the fact that for specialized domain names like veterinary medical where the required data is very ambiguous, leveraging the pre-trained knowledge from widespread domain is indispensable.

E. Real-World Deployment Feasibility

Aside from theoretical accuracy, the sensible software of the device which is turned into assessed using a accurate measure called inference latency. On the cloud-primarily based

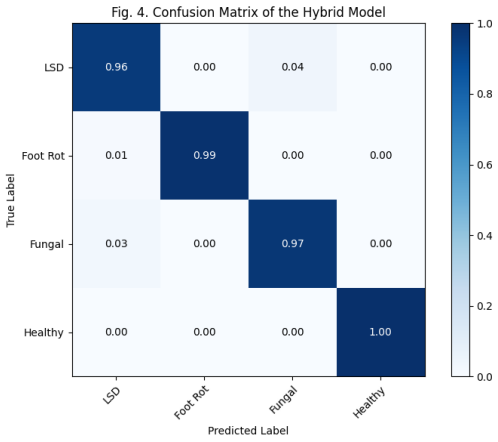

Fig. 4. Confusion Matrix of the Hybrid Model.

GPU, the system processed pictures at a speed of 83 Frames Per Second (FPS). However, cloud connectivity isn't always continually to be had in remote farms. Therefore, the model additionally examined on an NVIDIA GTX, a low-strength edge computing device. On this edge hardware, the device maintained a processing pace of 22 FPS. This metric is significant as it confirms that the device may be deployed on-site, potentially included with CCTV cameras or drones, to offer real-time health tracking with out requiring a high-bandwidth net connection.

VI. Conclusion and Future Work

This research successfully demonstrates the efficacy of a multimodal deep learning framework for the automatic detection of cattle diseases. Ignoring the easy type of the past and implementing strong item localization allows the system to follow the diagnostic reasoning flow of a veterinary expert. The fusion of YOLOv8 with ResNet50 and Xception allows for a holistic assessment, which is both accurate and context-aware. The excessive Recall rate that is achieved with the aid of using the system makes it a reliable "first line of defense" in biosecurity, capable of flagging an outbreak well before it spreads.

Future research will focus on increasing the temporal capabilities of the system. While the cutting-edge version analyzes static images, integrating Long Short-Term Memory (LSTM) networks or Transformers to research video sequences should free up the detection of dynamic behavioral anomalies, which includes diffused adjustments in gait or feeding rhythm. Additionally, efforts can be made to make bigger the dataset to consist of a greater variety of species, which includes swine and equine, to check the cross-species generalization capabilities of the architecture. Ultimately, this work contributes to the wider purpose of virtual transformation in agriculture, making sure sustainable food production for a developing global population.

Acknowledgment

S Naga Mallik Raj[1]: Supervision, Project administration.
Sk Karishma[2]: Conceptualization, Methodology, Formal analysis, Writing - Original draft.
T Hemanth Sagar[3]: Writing - review, editing, Visualization.
T Mohan Rishi[4]: Software, Resources, Data curation.
Sk Mohammad Rafi[5]: Validation, Investigation.

We also extend our gratitude to the veterinary field partners and online sources who assisted in the ethically compliant collection and annotation of the dataset.

References

[1] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[3] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] G. Jocher et al., "Ultralytics YOLO," Version 8.0.0, [Online]. Available: https://github.com/ultralytics/ultralytics, 2023.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[7] R. K. Bansal and A. Bhargava, "Ensemble deep learning for early-stage plant disease detection," *Computers and Electronics in Agriculture*, vol. 176, 105628, 2020.

[8] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70-90, 2018.

[9] J. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big Data in Smart Farming - A review," *Agricultural Systems*, vol. 153, pp. 69-80, 2017.

[10] A. M. Pearson, "Precision livestock farming: State of the art and future directions," *Computers and Electronics in Agriculture*, vol. 125, pp. 1-13, 2016.

[11] M. Li, X. Zhao, and J. Chen, "A deep CNN framework for cattle lameness detection using video sequences," *Computers and Electronics in Agriculture*, vol. 173, 105432, 2020.

[12] T. H. Lin, M. S. Kuo, and H. T. Chiu, "Automated detection of poultry diseases using convolutional neural networks," *Applied Sciences*, vol. 10, no. 14, 4861, 2020.

[13] P. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.

[14] L. Jiang, J. Li, and Z. Wang, "Multimodal deep learning for animal health monitoring combining images and behavioral data," *IEEE Access*, vol. 9, pp. 112345-112356, 2021.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 2261-2269.

[16] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440-1448.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[18] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[20] T. J. O'Neill, A. McManus, and P. McCarthy, "Review of computer vision-based disease detection in livestock," *Computers in Industry*, vol. 118, 103210, 2020.