International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Robust and explainable poultry disease classification via MaxViT with attention-guided visualization

Hasib Fardin [1], Hasan Md Imran [2, *], Hamdadur Rahman [3], Anamul Haque Sakib [4] and Md Ismail Hossain Siddiqui [5]

[1] *Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA.*
[2] *Department of Computer Science, The University of Alabama in Huntsville, Huntsville, AL 35899, USA.*
[3] *Department of Management Information System, International American University, Los Angeles, CA 90010, USA.*
[4] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA.*
[5] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1848-1859

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1165

Abstract

Poultry farming plays a crucial role in global food security, yet it faces significant challenges from infectious diseases such as Coccidiosis, Salmonella, and Newcastle disease, which can lead to substantial economic losses. Existing deep learning models often encounter issues like class imbalance, poor generalization across different datasets, and a lack of interpretability. To address these limitations, we propose a novel hybrid deep learning framework based on the MaxViT architecture. This framework combines MBConv blocks with both block-wise and grid-based self-attention mechanisms, allowing it to effectively capture local and global features in complex fecal images. In our study, we utilized two publicly available poultry fecal image datasets, consisting of 8,067 and 6,812 images, respectively. Each dataset includes four classes: Coccidiosis, Salmonella, Newcastle disease, and Healthy. To tackle the severe class imbalance, we applied data augmentation techniques. We evaluated the models using various metrics, including accuracy, F1-score, specificity, PR AUC, and MCC, under 10-fold cross-validation. Our proposed MaxViT model achieved impressive accuracy scores of 99.54% and 98.96% on the two datasets, outperforming ViT-B/16, ViT-L/32, DeiT-S, and T2T-ViT-14 across all metrics. Additionally, we integrated Grad-CAM to provide visual explanations of the model's decisions, thereby enhancing transparency and applicability in veterinary settings. This study introduces a deployable, interpretable, and highly accurate framework for intelligent poultry disease diagnosis, effectively addressing critical limitations found in previous research.

Keywords: Poultry disease; MaxViT; Livestock health; Veterinary diagnostics; Deep learning

1. Introduction

Poultry farming accounts for over 35% of global meat production and supports the livelihoods of approximately 1.3 billion people [1]. This sector particularly benefits smallholder farmers in low-income regions, where poultry can represent up to 60% of livestock income [2]. However, poultry diseases such as coccidiosis, salmonella, and Newcastle disease result in annual losses exceeding USD 3 billion [3]. These diseases undermine productivity, food safety, and public health. Traditional diagnostic methods, including manual inspections and laboratory tests, are often impractical in rural settings due to their high costs, time consumption, and the need for skilled personnel. Deep learning techniques, especially Convolution Neural Networks (CNNs) have improved automated disease detection by identifying local features like texture and color gradients [4], [5]. However, CNNs often struggle with understanding global spatial contexts and lack interpretability [6]. Vision Transformers (ViTs) provide a solution with their self-attention

[*] Corresponding author: Hasan Md Imran

mechanisms, which allow for the modeling of long-range dependencies and provide contextual analysis [7]. Nonetheless, ViTs require large datasets and do not possess the inductive biases of CNNs, which can limit their effectiveness in low-data or noisy agricultural environments.

While many studies have applied CNNs and early ViT models to poultry disease classification, several gaps remain. Few incorporate explainable AI (XAI) techniques, which limits interpretability. Additionally, no study has explored hybrid ViTs that combine CNN and transformer architectures for improved feature extraction. Critical issues such as class imbalance—especially for underrepresented diseases like Newcastle—have largely gone unaddressed, resulting in biased models. Furthermore, cross-dataset generalization and visual interpretability of predictions are seldom investigated, despite being essential for real-world deployment.

To address these challenges, this study has the following objectives:

- To design and implement a robust hybrid model that will capture both local and global visual patterns from fecal images.
- To develop a strategy to reduce class imbalance and enhance the model's robustness.
- To integrate XAI for improving interpretability and providing visual justifications for model predictions.
- To evaluate the model's performance across two diverse datasets and benchmark it against standard ViT variants.

To achieve our objectives, we propose a framework based on the MaxViT architecture, which was specifically selected for its ability to integrate the strengths of both convolutional and transformer-based models (See Figure 1). MaxViT uses depthwise convolutions, MBConv blocks, and both block-wise and grid-based attention mechanisms, allowing it to capture fine local details and broader spatial relationships in fecal imagery. This hybrid design helps identify subtle visual differences between disease types and enhances generalization in real-world conditions. We use two public datasets and apply targeted augmentation techniques to balance class distributions and minimize training bias. Further, we improve model interpretability with Grad-CAM visualizations, which show which regions in the input images affect predictions.

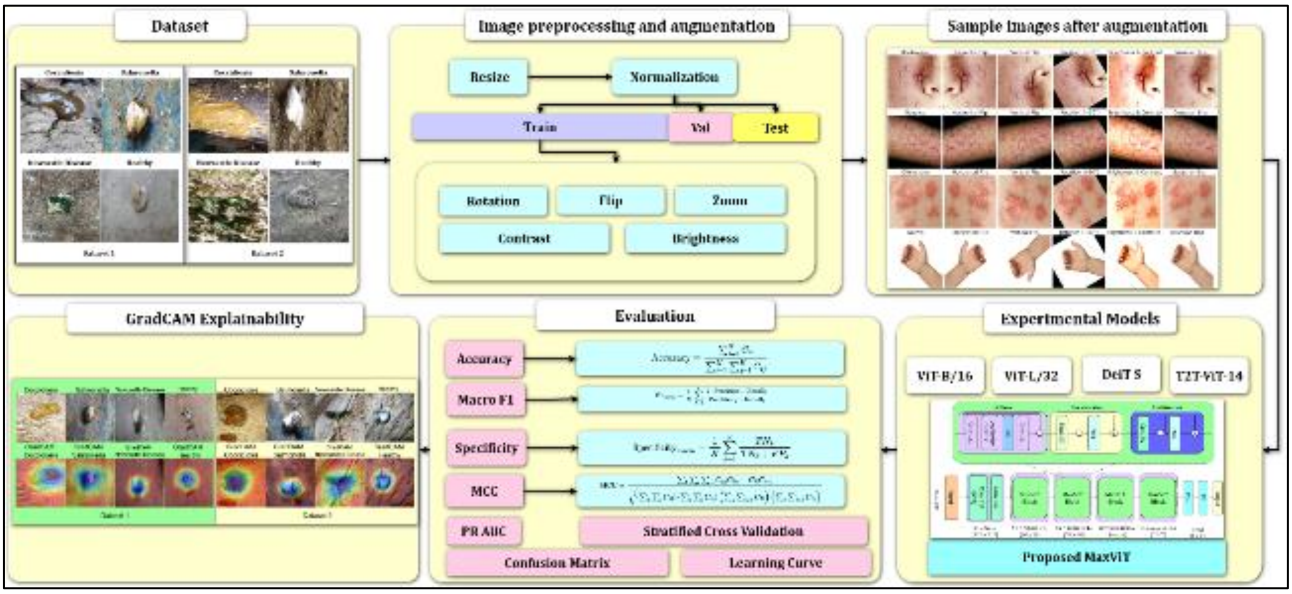

Figure 1 Overall methodology

Our key contributions are as follows:

- Proposed a new framework for classifying poultry diseases using the Multi-Axis ViT (MaxViT). It combines convolutional layers with block-wise and grid-based attention mechanisms to effectively capture local textures and global spatial dependencies.
- Addressed class imbalance in two poultry fecal image datasets by applying targeted data augmentation techniques. This improved model stability and performance for underrepresented classes like Newcastle Disease.
- Utilized Grad-CAM to provide visual interpretations of model predictions, improving trust and understanding in veterinary and agricultural applications.

- Conducted a comparative analysis of various models, demonstrating that our method surpasses existing approaches.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the datasets, preprocessing, and architecture. Section 4 presents the results and comparative analysis. Section 5 offers a critical discussion, and Section 6 concludes the study.

2. Related Works

In recent years, artificial intelligence techniques have played a vital role across a wide range of real-world applications, including sustainable agriculture [8], [9], [10], biodiversity systems [11], [12], and disease diagnosis [13], [14]. Pretrained CNN models have also shown promise for detecting diseases in poultry. Machuve et al. [15] developed a deep learning model to identify Coccidiosis, Salmonella, and Newcastle disease using 8,067 fecal images. Among various CNN architectures tested, MobileNetV2 achieved the highest accuracy at 98.02%. However, the system lacked integration with clinical metadata and real-time diagnostic capabilities, which limited its practical utility. To improve portability and enable real-time diagnosis, mobile-based solutions have emerged. Degu et al. [16] proposed a system that employs YOLOv3 for fecal image segmentation and ResNet50 for classifying diseases into four categories. Their model, trained on 10,500 images, achieved an accuracy of 98.7%. Despite this success, its reliance on a single imaging modality and the absence of broader clinical validation restricts its deployment in the field.

Hybrid learning strategies have also been employed to enhance classification accuracy. Luo et al. [17] built a transfer learning pipeline using 6,812 fecal images, combining nine pre-trained CNNs with traditional classifiers such as SVM, Logistic Regression, and KNN. The best-performing model, ResNet152-SVM, achieved an accuracy of 98.3%. However, the dataset was limited to chickens, affecting the model's generalizability to other poultry species. Studies have investigated how well models generalize across diverse environments. Chidziwisano et al. [18] evaluated models like MobileNet, DenseNet, and ResNet on datasets from Tanzania (8,077 images) and Malawi. While binary classification was effective across regions (MobileNet achieved 98% accuracy in Tanzania and 82% in Malawi), the performance in multiclass classification degraded significantly, revealing challenges related to breed differences, environmental factors, and dataset imbalance.

Efforts to enhance model interpretability have involved the use of ViTs. Luong et al. [19] developed a ViT16 model integrated with the Integrated Gradients technique to interpret predictions. Trained on 8,067 fecal images, it achieved 100% accuracy, outperforming both CNN and ViT32 models. However, its dependency on localized data and a single image modality may hinder its scalability to varied or real-time applications. Audio-based classification models have also been explored for the early detection of diseases. Cuan et al. [20] introduced a BiLSTM-based Deep Poultry Vocalization Network (DPVN) to detect Newcastle disease using chicken vocalizations. The model achieved an accuracy of 98.5% after advanced noise filtering and vocal segmentation. Its limitation lies in focusing on a single disease and the need for controlled acoustic environments. Lastly, lightweight hybrid architectures are gaining traction for efficient disease detection. He et al. [21] proposed ResTFG, a CNN-transformer hybrid model with 1.95 million parameters, designed to classify seven Eimeria species from 5,103 oocyst images. The model achieved an accuracy of 96.9% and a processing speed of 256 frames per second. However, it faced challenges in distinguishing morphologically similar species, such as E. praecox, indicating a need for improvement in fine-grained classification.

3. Materials and Methods

3.1. Data Description

This study uses two publicly available image datasets with labeled samples of poultry diseases. Each dataset includes four categories: Coccidiosis, Salmonella, Newcastle Disease, and Healthy. These categories represent common conditions that affect poultry health, making them essential for early detection and effective disease management. Dataset 1 [22] comprises a total of 8,067 samples, while Dataset 2 [23] contains 6,812 samples, resulting in a combined total of 14,879 images. In Dataset 1, the class distribution is relatively balanced among the Coccidiosis (2,476), Salmonella (2,625), and Healthy (2,404) classes. However, Newcastle Disease is significantly underrepresented, with only 562 samples. A similar pattern is observed in Dataset 2, where Coccidiosis (2,103), Salmonella (2,057), and Healthy (2,276) classes are well represented, while Newcastle Disease has only 376 samples. Sample image from both datasets illustrated in Figure 2.

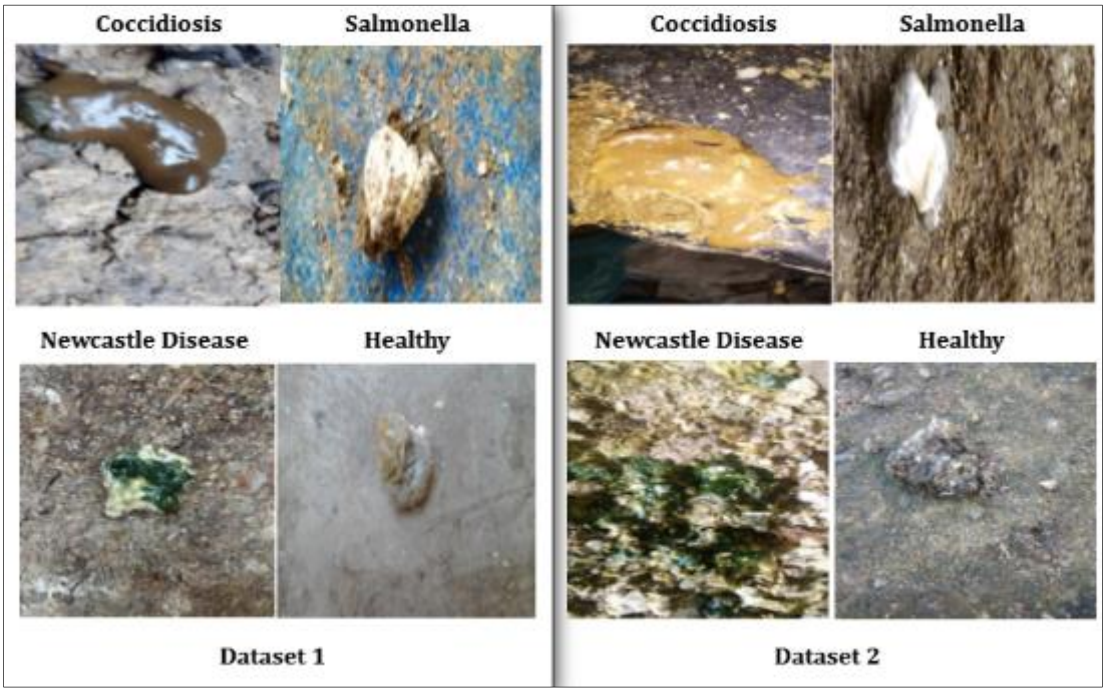

Figure 2 Sample image of each class from both datasets

3.2. Data Preprocessing and Augmentation

To prepare the poultry disease images for classification using the ViTs, several preprocessing techniques were applied to standardize image dimensions, enhance visual features, and normalize pixel intensity values. Initially, all images from the two poultry disease datasets were resized to 224×224 pixels to meet the input size requirements of the ViTs architecture. The pixel values were then normalized to a range of [0, 1] to ensure consistency in the input distribution [24]. After preprocessing, the datasets were divided into training, validation, and test sets using stratified sampling with a ratio of 70/10/20, which preserved the class distributions in each subset (See Table 1). This approach ensured that the model encountered a balanced representation of each disease class during evaluation.

Table 1 Class distribution after splitting

Dataset	Class	Total	Training (70%)	Validation (10%)	Testing (20%)
Dataset 1	Coccidiosis	2476	1733	247	496
	Salmonella	2625	1837	262	526
	Healthy	2404	1682	240	482
	Newcastle Disease	562	393	56	113
	Total	8067	5645	805	1617
Dataset 2	Coccidiosis	2103	1472	210	421
	Salmonella	2057	1439	205	413
	Healthy	2276	1593	227	456
	Newcastle Disease	376	263	37	76
	Total	6812	4767	679	1366

Both datasets initially showed class imbalance, particularly in the Newcastle Disease class, which contained significantly fewer samples than the others. To address this issue and enhance the model's generalization capabilities, data augmentation techniques were implemented. Common augmentation strategies included random rotations (±15°), horizontal and vertical flips, zooming, and adjustments to brightness and contrast. These augmentations were applied only to the training set and specifically targeted the underrepresented classes to synthetically increase their sample sizes. After augmentation, each class within the individual datasets was balanced to match the size of the majority class. In the first dataset, all classes were balanced to contain 1,837 images each, resulting in a total of 7,348 images. Similarly, in the second dataset, each class was augmented to have 1472 images, yielding a total of 5,888 images. This preprocessing and augmentation strategy ensured that the datasets were not only compatible with the model but also robust and balanced [25]. Figure 3 illustrates sample images from each disease after applying augmentation techniques.

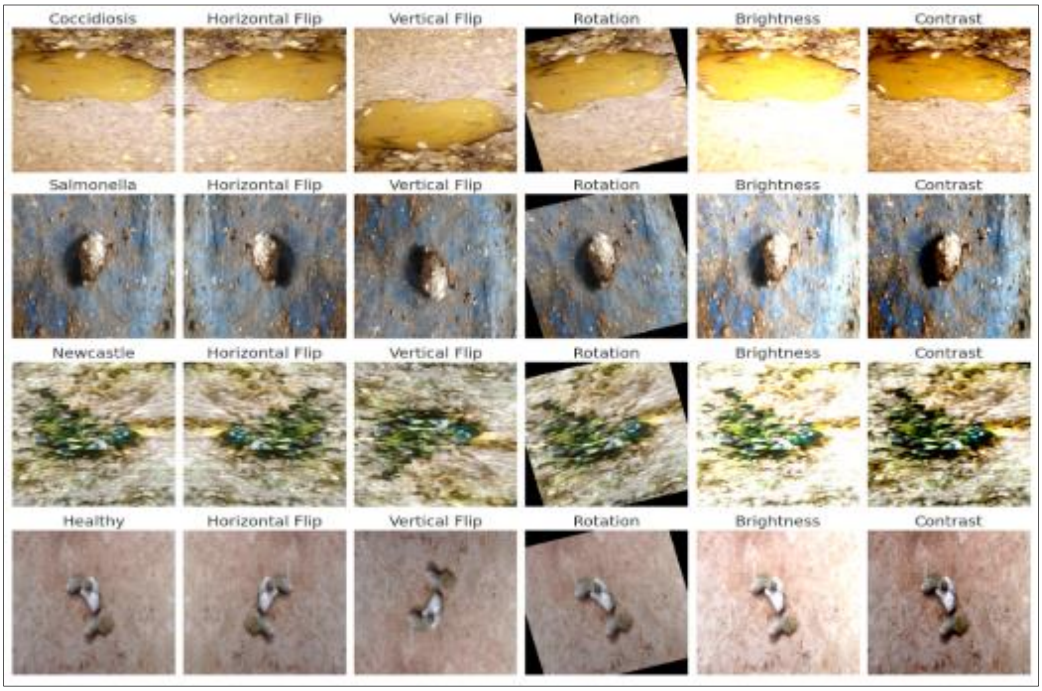

Figure 3 Sample augmented images for each poultry disease class

3.3. Experimental Models

3.3.1. ViT Models

In this study, we used five ViT models to classify poultry diseases from fecal images. These models were pretrained on the ImageNet dataset and fine-tuned to identify four specific health conditions in poultry. Unlike CNNs, ViTs utilize self-attention mechanisms to better capture subtle disease patterns that may be spread across the images. This makes them effective for our needs, as disease indicators are often diffused and require contextual understanding.

ViT-B/16, introduced by Dosovitskiy et al. [26], is an early transformer model designed for vision tasks. It processes images by dividing them into 16×16 patches and uses multi-head self-attention. We selected it as a baseline for its strong generalization and simplicity. ViT-L/32, a larger variation, uses 32×32 patches, enhancing global context understanding but sacrificing finer detail. We chose this model to explore the balance between parameter size and recognition performance, especially in distinguishing similar disease patterns. DeiT-S, developed by Facebook AI [27], serves as a lighter alternative to ViT. It is designed to train efficiently on smaller datasets without necessitating large-scale pretraining. This model incorporates a distillation token that benefits from knowledge transfer during training. Its lightweight design makes it suitable for deployment in resource-constrained environments while still maintaining competitive accuracy. T2T-ViT-14 improves upon the original ViT by refining the tokenization process [28]. Rather than simply flattening raw patches, it recursively aggregates neighboring tokens to better model local structures before applying self-attention. This hierarchical tokenization approach is particularly effective in disease classification tasks where localized features are critical.

3.3.2. Proposed MaxViT Model

This study aims to enhance poultry disease classification using the MaxViT architecture. MaxViT utilizes both CNN and transformer-based attention mechanisms, making it ideal for this task. It effectively captures local features, like texture and edge variations in poultry feces, while also understanding the global context needed to analyze complex patterns [29]. This dual capability allows it to handle the complexities of real-world poultry waste imagery. The overall architecture of MaxViT, as illustrated in Figure 4, starts with two convolutional layers. The first layer is a 3×3 convolution with a stride of 2, which performs spatial down sampling and basic edge detection. This is followed by another 3×3 convolution layer that further refines low-level features. These convolutional stages are essential for capturing early spatial cues and enhancing the model's inductive bias toward local structures.

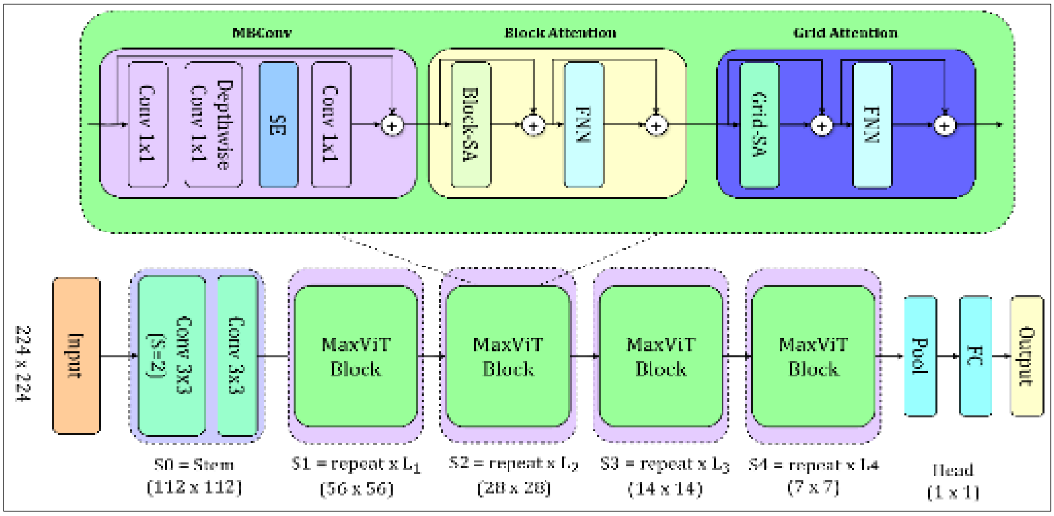

Figure 4 Proposed MaxViT architecture

The backbone of MaxViT consists of a series of MaxViT blocks, each featuring three key components: MBConv, Block Attention, and Grid Attention. The MBConv module, adapted from mobile networks, includes a 1×1 pointwise convolution, a depthwise convolution, and a Squeeze-and-Excitation (SE) block. This combination efficiently captures local patterns while maintaining computational efficiency [30]. Next, the Block Attention stage applies self-attention within non-overlapping blocks of the image, allowing the model to focus independently on localized regions. This is particularly useful for distinguishing detailed features such as speckled, foamy, or discolored textures associated with various diseases. To augment local attention, Grid Attention introduces a global perspective by enabling interactions across grid-partitioned sections of the image. This mechanism helps the model learn relationships between spatially distant regions, addressing the limitations of block attention and ensuring that contextual dependencies are preserved [31]. Following these attention stages, the architecture includes a global pooling layer, a fully connected layer, and a softmax-based classification layer that outputs predictions for the four target classes. MaxViT outperforms traditional transformers and CNN models, particularly in situations with subtle differences or noisy backgrounds.

3.4. Training Parameters and Evaluation

The training process for all models was conducted over 30 epochs using the Adam optimizer, with a constant learning rate set at 0.001. For the multiclass classification task, we employed categorical cross-entropy as the loss function. To prevent overfitting and improve the model's generalization ability, we integrated early stopping, a learning rate adjustment mechanism, and model checkpointing into the training workflow. Model performance was evaluated using five metrics: accuracy, F1 score, specificity, PR AUC, and MCC. Accuracy provided an overall measure of correct predictions, while the F1 score reflected the balance between precision and recall, which is particularly important when dealing with imbalanced classes. Specificity assessed the model's ability to correctly identify negative samples. PR AUC offered insights into the precision-recall relationship across various thresholds, and MCC provided a consolidated performance score by considering all elements of the confusion matrix. To ensure a reliable and fair evaluation of the models, we applied 10-fold stratified cross-validation. The dataset was divided into ten parts while preserving the original class distribution. In each fold, one subset was designated for validation, and the remaining nine were used for training. This process was repeated across all folds to ensure a comprehensive and unbiased evaluation.

4. Results and Discussion

Performance comparison of various ViT architectures on two imbalanced datasets, evaluated using five key metrics indicate that the proposed MaxViT model consistently outperforms all baseline models across both datasets. For Dataset 1, MaxViT achieves the highest performance across all metrics, with an accuracy of 98.34%. Similarly, for Dataset 2, MaxViT leads with a PR AUC of 98.22%. Among the other models, T2T-ViT-14 ranks second, demonstrating strong performance but consistently falling short of MaxViT. The DeiT-S model performs well on Dataset 1 but shows a significant decline on Dataset 2, particularly in the MCC metric (91.39%), which indicates challenges in generalization. Both ViT-L/32 and ViT-B/16 exhibit moderate performance, with ViT-B/16 generally receiving the lowest scores across most metrics. These variations highlight the architectural limitations of standard ViT variants when tackling imbalanced classification tasks. The ± values indicate standard deviations, which reflect model stability. MaxViT consistently demonstrates low variance (±0.29 in accuracy), suggesting reliable performance across different runs. In contrast, models such as ViT-B/16 show high variance (±1.58), making them less dependable. Overall, MaxViT exhibits both high accuracy and stability, making it particularly suitable for real-world applications that involve imbalanced data (Table 2).

Table 2 Performance comparison on imbalanced datasets

Dataset	Model	Accuracy	F1	Specificity	PR AUC	MCC
Dataset 1	Proposed MaxViT	98.34 ± 0.29	97.21 ± 0.33	98.03 ± 0.22	98.47 ± 0.38	96.84 ± 0.47
	T2T-ViT-14	96.87 ± 0.46	96.51 ± 0.41	96.75 ± 0.39	97.05 ± 0.33	95.91 ± 0.26
	DeiT-S	95.97 ± 0.59	95.68 ± 0.27	96.13 ± 0.45	95.62 ± 0.48	94.39 ± 0.42
	ViT-L/32	94.49 ± 0.31	94.02 ± 0.61	95.73 ± 0.61	95.11 ± 0.47	93.51 ± 1.16
	ViT-B/16	93.17 ± 1.13	93.53 ± 0.68	92.74 ± 0.93	94.00 ± 0.43	92.82 ± 1.06
Dataset 2	Proposed MaxViT	97.66 ± 0.44	96.75 ± 0.28	97.88 ± 0.49	98.22 ± 0.32	96.43 ± 0.53
	T2T-ViT-14	96.33 ± 0.68	94.79 ± 0.55	96.27 ± 0.63	95.89 ± 0.41	94.98 ± 0.39
	ViT-L/32	94.41 ± 0.56	95.09 ± 0.91	94.92 ± 0.46	95.13 ± 0.69	93.45 ± 0.41
	ViT-B/16	94.12 ± 1.58	95.39 ± 1.02	95.08 ± 1.18	95.18 ± 0.98	93.61 ± 1.23
	DeiT-S	91.82 ± 1.74	92.39 ± 1.19	92.55 ± 0.92	93.66 ± 1.07	91.39 ± 0.52

Table 3 Performance comparison on balanced datasets

Dataset	Model	Accuracy	F1 Score	Specificity	PR AUC	MCC
Dataset 1	Proposed MaxViT	99.54 ± 0.35	98.91 ± 0.47	99.13 ± 0.26	99.67 ± 0.41	97.04 ± 0.11
	T2T-ViT-14	98.67 ± 0.44	98.21 ± 0.48	97.95 ± 0.33	98.25 ± 0.39	97.11 ± 0.21
	DeiT-S	97.57 ± 0.61	97.28 ± 0.29	97.53 ± 0.46	96.82 ± 0.52	95.99 ± 0.45
	ViT-L/32	96.49 ± 0.36	96.02 ± 0.66	96.93 ± 0.65	96.31 ± 0.53	95.71 ± 1.22
	ViT-B/16	96.17 ± 1.21	96.13 ± 0.72	94.84 ± 0.97	96.20 ± 0.49	95.02 ± 0.62
Dataset 2	Proposed MaxViT	98.96 ± 0.39	98.45 ± 0.34	98.95 ± 0.53	99.22 ± 0.37	97.63 ± 0.58
	T2T-ViT-14	97.83 ± 0.71	96.49 ± 0.59	97.47 ± 0.61	98.09 ± 0.44	96.18 ± 0.42
	ViT-L/32	96.41 ± 0.63	96.69 ± 0.95	96.12 ± 0.49	97.33 ± 0.74	95.65 ± 0.44
	ViT-B/16	96.12 ± 0.61	96.89 ± 0.87	96.28 ± 0.56	97.38 ± 0.78	95.81 ± 0.42
	DeiT-S	93.92 ± 1.81	94.49 ± 1.24	94.75 ± 0.74	95.86 ± 1.13	93.59 ± 0.55

Table 3 summarizes the performance of various ViT-based models on two balanced datasets. The proposed MaxViT model achieves the highest scores on both datasets. For Dataset 1, MaxViT records a top accuracy of 99.54% and a specificity of 99.13%. Similarly, on Dataset 2, MaxViT leads with an F1 score of 98.45% ± 0.34, demonstrating its strong generalization on well-balanced data. In second place overall is T2T-ViT-14, which performs closely behind MaxViT, particularly in Dataset 1, with a Matthews Correlation Coefficient (MCC) of 97.11%. The DeiT-S and ViT-B/16 models also perform reasonably well, but their metrics exhibit greater variation in Dataset 2. For instance, DeiT-S shows a drop in accuracy to 93.92%, indicating reduced consistency. ViT-L/32 generally maintains moderate performance across both datasets. The standard deviation values indicate that MaxViT consistently shows the lowest deviation across most metrics (e.g., ±0.11 MCC in Dataset 1), suggesting reliable and reproducible results across different runs. In contrast, DeiT-S and ViT-L/32 exhibit higher variability, especially on Dataset 2, with some metrics exceeding ±1.0, which indicates less stable behavior during repeated training.

Figure 5 demonstrates strong and stable model performance across all evaluated metrics over the course of 30 training epochs for both Dataset 1 and Dataset 2. For Dataset 1, both the training and validation loss curves exhibit a consistent downward trend, reaching near-zero values by the end of training. While minor spikes are observed in the validation loss between epochs 5 and 15, the curves eventually stabilize, indicating effective convergence without signs of overfitting. The accuracy improves rapidly within the first 10 epochs, surpassing 95% and stabilizing between 98% and 99% for both training and validation sets, with only a minimal gap suggesting strong generalization. Similarly, the recall and precision curves show an upward trend, approaching 98% by the final epochs, indicating reliable true positive detection and confident classification. In contrast, Dataset 2 demonstrates a more stable training pattern overall. The loss curves decline smoothly and consistently, with values falling below 0.05, reflecting robust optimization. Accuracy increases quickly and remains above 98% after approximately 15 epochs, with the training and validation curves staying closely aligned. Likewise, recall and precision improve steadily and plateau near 98%, confirming the model's strong and consistent classification performance across both training and validation sets.

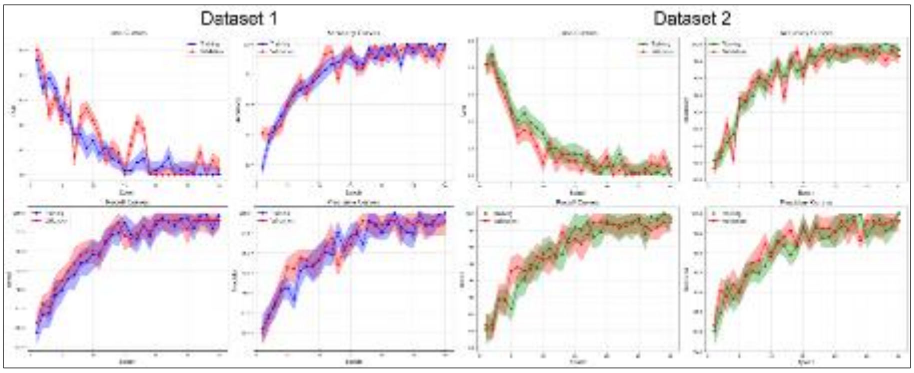

Figure 5 Learning curve of the proposed MaxViT model

Figure 6 presents the confusion matrices for the proposed MaxViT model, evaluated on two different datasets. On Dataset 1, the model achieved a high accuracy of 99.54%, correctly classifying nearly all samples across four classes: 493 out of 496 for Coccidiosis, 523 out of 528 for Salmonella, 479 out of 480 for Healthy, and 112 out of 113 for Newcastle Disease. The low number of misclassifications indicates a strong understanding of inter-class boundaries. In Dataset 2, the model attained an accuracy of 98.96%, with correct classifications of 416 out of 420 for Coccidiosis, 408 out of 413 for Salmonella, 450 out of 455 for Healthy, and 75 out of 78 for Newcastle Disease. Although Dataset 2 exhibited a slightly higher misclassification rate, particularly among visually similar classes, the model's performance remained consistently robust, demonstrating its strong ability to generalize.

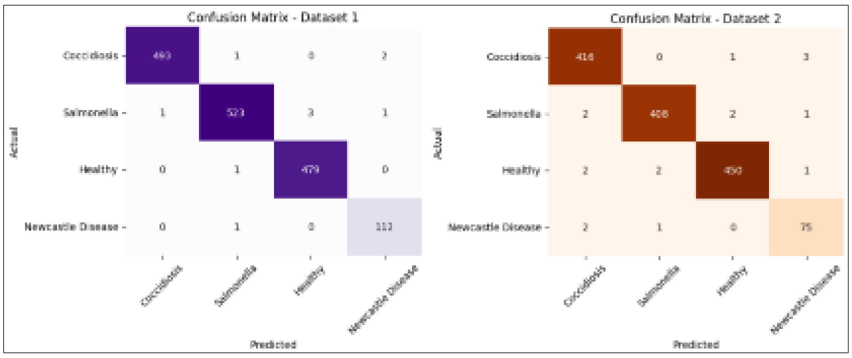

Figure 6 Confusion matrix of the proposed MaxViT model

The Grad-CAM visualization in Figure 7 demonstrates our proposed model's interpretability, trained on Dataset 1 (green background) and Dataset 2 (yellow background). The top row features raw images of poultry droppings, while the bottom row presents Grad-CAM heatmaps highlighting the key areas influencing the model's predictions. In Dataset 1, the model identifies key features for each disease class. For Coccidiosis and Newcastle Disease, the activation maps highlight discolored and watery droppings, indicating infection. For Salmonella, it emphasizes the solid central mass of feces. The heatmap for the Healthy class shows a uniform coverage of well-formed droppings. However, variations in background and lighting have slightly impacted the consistency of focus across samples. In Dataset 2, the model exhibits strong, symmetric activations across all disease classes, demonstrating its effectiveness in feature identification. The heatmaps for Coccidiosis and Newcastle Disease highlight distinct central areas, while Salmonella and Healthy samples show clear regions of interest. The improved textures and consistent backgrounds in Dataset 2 enhance clarity and feature localization.

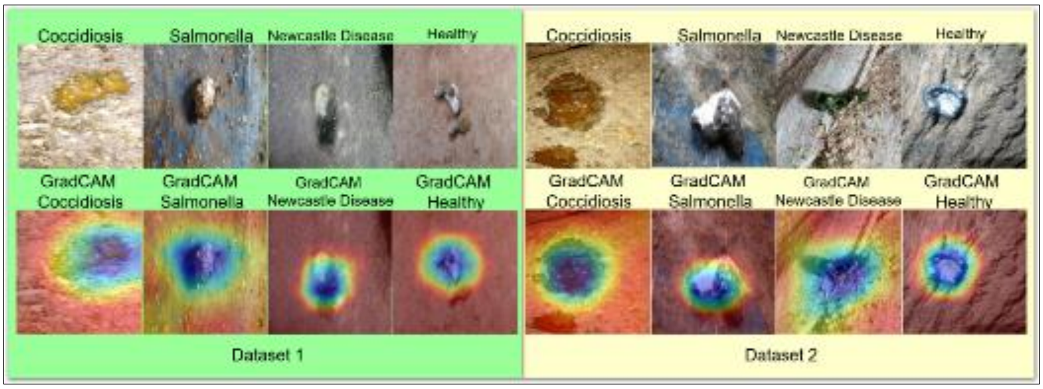

Figure 7 Sample Grad-CAM outputs showing key regions identified by the MaxViT model

Table 4 presents the outstanding performance of the proposed MaxViT models, which achieved accuracies of 99.54% and 98.96% on two different datasets. Unlike previous models, MaxViT incorporates XAI through Grad-CAM, which enhances both interpretability and reliability for real-world poultry disease diagnostics. Among prior approaches, YOLOv3 combined with ResNet50 and BiLSTM models showed strong performances with accuracies of 98.70% and 98.50%, respectively. However, these models lack XAI integration and are limited in their modality or disease coverage. While MobileNetV2 and ResNet152 combined with SVM also performed well, they struggled to generalize across different regions or poultry types, which reduced their robustness. In contrast, ResTFG and Deep CNN approaches demonstrated lower accuracy and lacked transparency, making them less suitable for practical deployment. MaxViT provides a robust and interpretable solution for intelligent poultry disease monitoring by addressing critical limitations found in previous methods.

Table 4 Comparison of existing models with the proposed MaxViT

Ref. No.	Model	Data	Result	XAI
Ours	Poposed MaxViT	8067	99.54%	Yes
	Proposed MaxViT	6812	98.96%	Yes
[15]	MobileNetV2	8067	98.02%	No
[16]	YOLOv3 + ResNet50	10500	98.70%	No
[17]	ResNet152 + SVM	6812	98.30%	No
[18]	MobileNet	8077	98% (Tanzania)	No
[19]	ResTFG (CNN + Transformer)	5103	96.90%	No
[20]	BiLSTM	8000	98.50%	No
[21]	Deep CNN	Not specified	91.47%	No

The MaxViT model outperforms others due to its hybrid architecture that captures both local and global features. Its MBConv blocks are designed to extract fine details, like discoloration and structural irregularities in fecal images. The block-wise and grid-based self-attention mechanisms enhance spatial reasoning across different image regions. This multi-axis attention structure effectively addresses the limitations of CNNs in managing long-range dependencies and the challenges faced by ViTs in low-data scenarios. The low standard deviations in performance metrics indicate the model's stability and generalization, crucial for real-world applications. Data augmentation improves model robustness, especially for underrepresented classes like Newcastle Disease. The targeted augmentation approach increased intra-class diversity and reduced biased learning. The improvements in F1-score and MCC demonstrate its effectiveness in reducing false negatives and enhancing classification reliability for all disease categories. Grad-CAM enhances interpretability by highlighting important areas in images, such as discolored textures in Coccidiosis and dense central masses in Salmonella. This transparency allows domain experts to validate the model’s clinically relevant decisions, fostering trust and usability in veterinary settings. Additionally, Grad-CAM's visual feedback is crucial for making explainable decisions in disease intervention. The model is ideal for low-resource agricultural settings. MaxViT can be fine-tuned using small, domain-specific datasets and comes with preloaded ImageNet weights, allowing for scalability. It can be further optimized with techniques like quantization or pruning to deploy on mobile or edge devices, enabling quick on-site disease diagnosis without needing centralized labs or specialist knowledge.

Despite these strengths, the study has several technical limitations. The augmentation process, while helpful, can create distributional shifts that impact generalization in real-world situations. MaxViT's grid attention is computationally intensive, which may limit real-time use on edge devices. Moreover, the model only uses image data and does not incorporate contextual information—like breed type or environmental metadata—that could enhance its diagnostic accuracy, especially in ambiguous cases. Future work should prioritize creating lightweight MaxViT variants with linear attention to lower training and inference complexities. Incorporating multi-modal data, such as fecal images, audio signals, and environmental metadata, could enhance diagnostic accuracy. Using self-supervised contrastive learning on unlabeled poultry images may improve adaptability. Lastly, adding spatio-temporal models would enable tracking of disease progression, aiding early intervention and predictive analytics in poultry health management.

5. Conclusion

This study introduces a novel and robust framework for classifying poultry diseases using the MaxViT architecture. This architecture uniquely combines convolutional layers with block-wise and grid-based self-attention mechanisms. Unlike existing models, our approach effectively captures both local textures and global spatial dependencies in complex fecal images, leading to superior performance across multiple evaluation metrics. A key contribution of our work is a targeted augmentation strategy that addresses class imbalance, particularly improving representation for under-sampled diseases like Newcastle disease. Additionally, the integration of Grad-CAM enhances model transparency, providing essential visual interpretability for practical use in veterinary diagnostics. Extensive experiments conducted on two diverse datasets confirm the model's stability, accuracy, and generalizability. By addressing key limitations of existing CNN and ViT approaches—such as poor explainability, inadequate generalization, and class imbalance—this research marks a significant advance toward deployable, AI-driven systems for poultry disease monitoring. Future research will focus on lightweight deployment and the incorporation of multimodal diagnostic inputs.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] S. Bhimrao Khade, R. S. Khillare, and M. B. Dastagiri, "Global livestock development: Policies and vision," Indian Journal of Animal Sciences, vol. 91, no. 9, pp. 770–779, 2021, doi: 10.56093/ijans.v91i9.116470.

[2] D. Chen, K. Mechlowitz, X. Li, N. Schaefer, A. H. Havelaar, and S. L. McKune, "Benefits and Risks of Smallholder Livestock Production on Child Nutrition in Low- and Middle-Income Countries," Front Nutr, vol. 8, p. 751686, Oct. 2021, doi: 10.3389/FNUT.2021.751686/XML/NLM.

[3] E. Abebe and G. Gugsa, "A Review on poultry coccidiosis," Abyssinia Journal of Science and Technology, vol. 3, no. 1, pp. 1–12, Jun. 2018, Accessed: Apr. 14, 2025. [Online]. Available: https://www.ajol.info/index.php/abjst/article/view/281032

[4] A. Al Noman et al., "Monkeypox Lesion Classification: A Transfer Learning Approach for Early Diagnosis and Intervention," Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024, pp. 247–254, 2024, doi: 10.1109/IC3I61595.2024.10828678.

[5] R. Haque et al., "Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning," BioMedInformatics 2024, Vol. 4, Pages 966-991, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[6] R. Haque, P. B.D, M. K. Hasan, A. H. Sakib, A. U. Rahman, and M. B. Islam, "Scientific Article Classification: Harnessing Hybrid Deep Learning Models for Knowledge Discovery," 2023 International Conference on Ambient

Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), pp. 1–7, Nov. 2023, doi: 10.1109/AIKIIE60097.2023.10389945.

[7] D. Wu et al., "Feature First: Advancing Image-Text Retrieval Through Improved Visual Features," IEEE Trans Multimedia, vol. 26, pp. 3827–3841, 2024, doi: 10.1109/TMM.2023.3316077.

[8] Md. R. Ahmed et al., "Towards Automated Detection of Tomato Leaf Diseases," 2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT), pp. 387–392, May 2024, doi: 10.1109/ICEEICT62016.2024.10534559.

[9] R. Haque et al., "Advancements in Jute Leaf Disease Detection: A Comprehensive Study Utilizing Machine Learning and Deep Learning Techniques," 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON), pp. 248–253, Sep. 2024, doi: 10.1109/PEEIACON63629.2024.10800378.

[10] M. S. Rahman et al., "Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models," 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 59–64, 2024, doi: 10.1109/RAAICON64172.2024.10928353.

[11] M. D. Hosen et al., "Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning," 2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[12] R. B. Cooper, J. T. Flannery-Sutherland, and D. Silvestro, "DeepDive: estimating global biodiversity patterns through time using deep learning," Nature Communications 2024 15:1, vol. 15, no. 1, pp. 1–14, May 2024, doi: 10.1038/s41467-024-48434-7.

[13] R. Muthukrishnan, S. Kannan, R. Prabhu, Y. Zhao, P. Bhowmick, and M. J. Hasan, "Tracking and Estimation of Surgical Instrument Position and Angle in Surgical Robot Using Vision System," 2023 International Conference on Network, Multimedia and Information Technology, NMITCON 2023, 2023, doi: 10.1109/NMITCON58196.2023.10275983.

[14] P. Podder, F. B. Alam, M. R. H. Mondal, M. J. Hasan, A. Rohan, and S. Bharati, "Rethinking Densely Connected Convolutional Networks for Diagnosing Infectious Diseases," Computers 2023, Vol. 12, Page 95, vol. 12, no. 5, p. 95, May 2023, doi: 10.3390/COMPUTERS12050095.

[15] D. Machuve, E. Nwankwo, N. Mduma, and J. Mbelwa, "Poultry diseases diagnostics models using deep learning," Front Artif Intell, vol. 5, p. 733345, Aug. 2022, doi: 10.3389/FRAI.2022.733345/BIBTEX.

[16] M. Z. Degu and G. L. Simegn, "Smartphone based detection and classification of poultry diseases from chicken fecal images using deep learning techniques," Smart Agricultural Technology, vol. 4, p. 100221, Aug. 2023, doi: 10.1016/J.ATECH.2023.100221.

[17] Y. Luo, Y. Chen, and A. P. P. Abdul Majeed, "Optimizing poultry disease classification: A feature-based transfer learning approach," Smart Agricultural Technology, vol. 10, p. 100856, Mar. 2025, doi: 10.1016/J.ATECH.2025.100856.

[18] G. Chidziwisano, E. Samikwa, and C. Daka, "Deep learning methods for poultry disease prediction using images," Comput Electron Agric, vol. 230, p. 109765, Mar. 2025, doi: 10.1016/J.COMPAG.2024.109765.

[19] H. H. Luong and T. M. Nguyen, "Improving Chicken Disease Classification Based on Vision Transformer and Combine with Integrated Gradients Explanation," International Journal of Advanced Computer Science and Applications, vol. 15, no. 4, p. 1235, Apr. 2024, doi: 10.14569/IJACSA.2024.01504124.

[20] K. Cuan, T. Zhang, Z. Li, J. Huang, Y. Ding, and C. Fang, "Automatic Newcastle disease detection using sound technology and deep learning method," Comput Electron Agric, vol. 194, p. 106740, Mar. 2022, doi: 10.1016/J.COMPAG.2022.106740.

[21] P. He et al., "A reliable and low-cost deep learning model integrating convolutional neural network and transformer structure for fine-grained classification of chicken Eimeria species," Poult Sci, vol. 102, no. 3, p. 102459, Mar. 2023, doi: 10.1016/J.PSJ.2022.102459.

[22] "Chicken Disease Image Classification." Accessed: Apr. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/allandclive/chicken-disease-1?resource=download&select=Train

[23] "Poultry Diseases Detection." Accessed: Apr. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/kausthubkannan/poultry-diseases-detection?select=poultry_diseases

[24] A. Al-Sakib, F. Islam, R. Haque, M. B. Islam, A. Siddiqua, and M. M. Rahman, “Classroom Activity Classification with Deep Learning,” 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, 2024, doi: 10.1109/ICICACS60521.2024.10498187.

[25] R. Haque et al., “A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis,” 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 255–262, Sep. 2024, doi: 10.1109/IC3I61595.2024.10829132.

[26] A. Dosovitskiy et al., “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE”, Accessed: Apr. 13, 2025. [Online]. Available: https://github.com/

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” Jul. 01, 2021, PMLR. Accessed: Apr. 14, 2025. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[28] L. Yuan et al., “Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet,” 2021. Accessed: Apr. 14, 2025. [Online]. Available: https://github.com/yitu-opensource/T2T-ViT

[29] Y. Chen, “A novel image classification framework based on variational quantum algorithms,” Quantum Inf Process, vol. 23, no. 10, pp. 1–28, Oct. 2024, doi: 10.1007/S11128-024-04566-9/FIGURES/11.

[30] X. Fu, R. Lin, W. Du, A. Tavares, and Y. Liang, “Explainable hybrid transformer for multi-classification of lung disease using chest X-rays,” Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–19, Feb. 2025, doi: 10.1038/s41598-025-90607-x.

[31] M. Sarıateş and E. Özbay, “A Classifier Model Using Fine-Tuned Convolutional Neural Network and Transfer Learning Approaches for Prostate Cancer Detection,” Applied Sciences 2025, Vol. 15, Page 225, vol. 15, no. 1, p. 225, Dec. 2024, doi: 10.3390/APP15010225.